heg

Haute école de gestion
de Neuchâtel

ILCE Institut de lutte contre la criminalité économique
Etudes postgrades HES

Travail de diplôme

TITRE : Maîtriser la Société de l'Information : Veille stratégique, détection de signaux faibles, due diligence et recherche d'informations en vue de la mise en place de systèmes de prévention contre les nouveaux risques liés la criminalité économique et au blanchiment d'argent.

Nom et prénom de l'étudiant : Koch Stéphane

Clause de confidentialité : ☐ oui X non

Date de remise du travail : 15 février 2003

Table des matières

1	**Introduction**	**1**
2	**Société de l'information, une nouvelle donne pour les entreprises connectées**	**2**
3	**Nouvelles technologies de l'information et de la communication (NTIC) : quels impacts pour les entreprises ?**	**3**
4	L'information aujourd'hui : Internet et le Web, catégorisation et définition	5
5	**L'aspect multicouche de l'information**	**6**
	5.1) L'aspect multicouche de l'information: Le Web	7
	5.2) L'aspect multicouche de l'information, l'adresse IP	7
	5.3) Exemple de redirection d'un site Web dans le but de commettre une fraude	8
6	**La recherche d'information sur le WEB**	**9**
	6.1) Les principales sources d'information	9
	6.2) Cartographie de l'information disponible sur le Web	9
	6.3) Les types d'outils et leurs fonctions	10
	6.4) Les différents outils de recherches de l'information	11
	6.5) Autres outils de recherches et traitement de l'information	12
	6.6) La définition des zones et périmètres de recherches	12
	6.7) Principes de bases pour établir une recherche	13
	6.8) Utilisation des fonctionnalités avancées des moteurs de recherches : les principaux « Opérateurs Booléens » :	14
7	**L'analyse et la crédibilité de l'information (inclus traçabilité et identification)**	**14**
	7.1) Crédibilité de l'information, présentation de deux cas école	15
	7.2) Crédibilité de l'information : les démarches de validation « logiques »	16
	7.3) Crédibilité de l'information : les démarches de validation « techniques »	16
	7.4) Remarque concernant l'identification des détenteur d'un site Internet	18
8	**Analyse de l'environnement et de la survenance de l'information par l'interprétation des signaux faibles**	**19**
	8.1) Modèle d'analyse de l'environnement de l'information	19
	8.2) Deuxième axe d'analyse du fait	19
	8.3) Utilité et compréhension du modèle d'analyse des signaux faibles	20
	8.4) La méthode « PUZZLE » d'analyse des signaux faibles	20
	8.5) Traitement de l'information : Principes de l'intelligence collaborative	23
	8.6) Exemples d'utilisation des méthodes traités dans le chapitre 8	23
9	**Stratégies visant à la maîtrise des Flux informationnels**	**24**
	9.1) La veille stratégique: principes de base et possibilités d'utilisation	24

	9.2) Philosophie pour la mise en place d'une structure de veille	25
	9.3) Etapes du cadre méthodologique d’une surveillance électronique:	25
	9.4) Actions défensives et préventives	25
10	**Conlusions**	**27**
I	Annexe : Glossaire des Termes Internet	
II	ANNEXE II Noms de domaine et glossaire des Acronymes	
III	Annexe : TEXT MINING ET INTELLIGENCE ECONOMIQUE : AUJOURD’HUI ET DEMAIN	
IV	ANNEXE 1V , La répartition géo-stratégique de l’Internet	

1) Introduction :

Objectif du travail :

Offrir des outils ainsi que les bases d'une méthodologie de prévention et de gestion des risques (liés à la criminalité économique et au risque de blanchiment d'argent) par à une perception et une compréhension des enjeux de la société de l'information - que l'on pourrait définir par une modélisation et une formalisation des courants tacites[1] de la société actuelle vers l'Internet, et les conséquences induites par l'utilisation des nouvelles technologies pour ce qui touche à la criminalité économique et celles, indirectes, liées au blanchiment d'argent. Une connaissance et une utilisation adéquate des moyens disponibles pour la gestion de son environnement informationnel, pourront permettre aux différents acteurs d'appréhender : le risque technologique et humain, les méthodes de recherches d'information et la « consistance » même de celle-ci, l'identification des différents intermédiaires et propriétaires de sites Internet, les flux informationnels présents dans l'environnement des entreprises, le risque à l'image, l'usurpation de d'identité, la traçabilité géographique des informations.

Publics visés et considérations d'ordre générales :

Ce document s'adresse plus particulièrement aux petites et moyennes entreprises. Pour les structures plus importantes il pourra servir de base de réflexion à l'élaboration de la stratégie de gestion de l'information et du risque dans les départements addoc (management, compliance, communication). Les aspects d'ordre techniques abordés dans ce document sont considérés comme les bases indispensables à la compréhension et à la gestion des problèmes abordés. Les différents termes techniques ainsi que les acronymes figurants dans ce mémoire sont documentés dans les annexes mentionnées dans la table des matières.

La société de l'information ne peut être abordée sans une connaissance de ses principales composantes, à cet effet, ce document traitera de trois volets essentiels à la compréhension des répercussions de l'utilisation des NTIC (Nouvelles technologie de l'information et de la communication). Le premier volet parlera plus spécifiquement des aspects liés aux systèmes d'information (SI) et à leur interconnexion. Le deuxième volet sera consacré à la consistance même de l'information numérisée, son analyse, sa crédibilité et aux diverses possibilités de recherches et d'identification et d'interprétation des données présentes au sein des sources disponibles sur le Net. Pour terminer, ce document traitera de la maîtrise des flux informationnels et la mise en place d'une structure préventive de gestion et de surveillance de « l'infosphère ». L'anticipation des risques à l'information, les méthodes défensives et mesures d'urgences pour lutter contre des « frappes informationnelles[2] ». Cette notion d'anticipation est décisive dans les domaines liés à la prévention de la criminalité économique et du risque de blanchiment d'argent. Avec l'émergence de la société numérique, il est devenu vital d'avoir la capacité de surveiller un environnement de plus en plus complexe, ainsi que de développer des méthodologies de recherche, d'identification et de traitement de l'information plus performante.

Pour arriver à ce résultat, il est donc nécessaire de connaître les différents outils et sources d'information disponibles sur le Web. Ces connaissances pourront ainsi permettre d'identifier, entre autres, les ayants-droits économiques de sites Web, de trouver de l'information sur les personnages politiquement exposés et de détecter, de manière anticipative, les risques inhérents à une situation donnée. La notion d'intelligence économique dans le traitement de l'information ainsi collectée, pourra permettre de produire un support d'aide à la décision pour la prise en compte des problèmes de criminalité économique et de la gestion du risque au niveau du traitement des flux financiers (blanchiment d'argent).

1)On pourrait aborder le courant tacite de la société comme un comportement ou une connaissance, qui ne provient pas directement d'un apprentissage mais d'une relation entre des connaissances acquises et l'exercice de celles-ci, ce qui génèrerait le développement d'un nouveau savoir (qui n'est inscrit nulle part, qui n'est pas présent sur un quelconque support, mode de croyances sociales)

2) Terme dont la paternité revient à Chritian Harbulot, Directeur de l'Ecole de Guerre Economique – www.ege.eslsca.fr

2) Société de l'information, une nouvelle donne pour les entreprises connectées :

La première conséquence de l'utilisation des ordinateurs et de l'Internet est la nécessité d'une dématérialisation de l'information amenée par la numérisation des échanges. La typologie même du réseau mondial a apporté des changements majeurs dans la définition du périmètre de l'entreprise. Les frontières géographiques classiques ont laissé la place à des territoires virtuels dont la délimitation – plus floue - peut dès lors se représenter en termes de segments de marchés et de secteurs d'activité. L'interconnexion croissante de la sphère professionnelle et la vitesse de propagation des (NTIC) ont généré de nouveaux risques pour les entreprises utilisant des SI.

Les développements actuels de la société de l'information ont rendu les aspects sociologiques indissociables des aspects technologiques, créant une interdépendance entre des spécifications d'ordre purement techniques et leurs répercussions sur notre mode de société. Cela à une influence sur notre réaction envers les SI, et ceux-ci sont tributaires de notre comportement dans leur mode de fonctionnement. L'écart intellectuel qui se créer de jour en jour au niveau de la « compréhension des impacts d'un SI dans une entreprise vis-à-vis de son secteur d'activité professionnel » est dû à la vitesse de l'avancée technologique et à la multiplication des facteurs de risques induits par les nouvelles méthodes de traitement de l'information. Il en résulte que les dirigeants d'entreprises continuent de penser que les machines connectées en réseau sont encore des outils de travail que l'on peut dissocier de la vision stratégique du fonctionnement de l'entreprise.

Source : www. mi2g.com /

Le graphique ci-dessus démontre on ne peut mieux la palette des risques inhérents à l'utilisation d'un SI. La notion de dépendance technologique revêt une importance considérable dans le présent document

3) Nouvelles technologies de l'information et de la communication (NTIC) : quels impacts pour les entreprises ?

Typologie des cinq catégories des risques liés aux NTIC:

Le tableau ci-dessous démontre les relations directes entre l'utilisation de systèmes d'information et l'activité même de l'entreprise :

	2001		2002	
	Pertes déclarées	%	Pertes déclarées	%
Vol d'information	161M$	43%	189 M$	42%
Fraude économique	102M$	27%	116 M$	25%
Intrusion	60 M$	16%	68 M$	15%
Sabotage	55 M$	14%	83 M$	18%
Total	**378 M$**	**100%**	**456 M$**	**100%**

Impact économique des attaques subies – source : CSI / FBI 2002

La couche matérielle (matériel) : le manque de fiabilité du matériel et ses diverses possibilités de disfonctionnements (coupure électrique, dommage matériel, malveillance, disfonctionnement) apportent une contrainte supplémentaire dans la gestion stratégique de l'entreprise. Cet aspect du problème – d'importance vitale – est souvent négligé en raison des surcoûts qu'il entraîne au niveau de la maintenance et de la mise en place de solution de secours (back up matériel et logiciel, sécurisation de la continuité de l'approvisionnement en électricité, protection des locaux contre le vol et le feu).

Le risque logiciel : Les modèles économiques actuels poussent les fabricants de logiciels à une course à la nouveauté, ce qui les poussent à diffuser des produits inaboutis sur le marché. La plupart des programmes disponibles à ce jour comportent un nombre élevé de possibilités de disfonctionnement ainsi que de sérieux manquements au niveau de la sécurité. En dehors de l'aspect éthique de la question, cela représente un risque de fuite d'information, d'espionnage industriel ou de perte de données (voir le tableau : « *Impact économique des attaques subies* »). La complexité même des programmes utilisés par les entreprises « autorise » aussi des utilisations détournées de ceux-ci par des personnes au bénéfice d'une éducation technique de base. Des facteurs extérieurs telle que la nécessité constante de mises à jour du système par le biais d'Internet, avec des données qui ne peuvent pas être contrôlées représentent des risques de discontinuités de fonctionnement supplémentaires (ou un risque stratégique pour les états ou pour les sociétés multinationales).

Le risque réseau : L'interconnexion des SI a permis non seulement une augmentation de la vitesse des échanges mais aussi du volume de ceux-ci. La capacité des SI à traiter un grand nombre de données, ainsi que la possibilité de traiter des informations de type hétérogènes de manière délocalisée, à modifié non seulement le comportement de l'entreprise, mais aussi le type de données qui transitent au travers les réseaux informatiques. En effet, pour rester concurrentielles et profiter pleinement des capacités offertes par les SI, on a commencé à formaliser (transformé en un format numérique compréhensible par les machines) le savoir tacite (la connaissance stratégique de l'entreprise). Ces données, qui classiquement étaient difficiles d'accès car elles se trouvaient réparties dans différents dossiers et dans les cerveaux des dirigeants, ont été regroupées et standardisées dans un langage d'échanges unique (TCP/IP).

Le risque humain : La complexité des SI a provoqué la nécessité d'accéder à des compétences différentes de celles en relation avec le secteur correspondant à l'activité professionnelle des entreprises concernées. L'évaluation même des connaissances nécessaires à la gestion et à la

maintenance d'un SI en est rendue que plus difficile. Cette situation et le manque de compréhension de l'importance des données qui transitent au sein du réseau ont créé, dans la plupart des entreprises, la croyance que la gestion du SI est uniquement un problème d'ordre technique. Le risque « humain » a été négligé par des dirigeants qui ont donné des pouvoirs exceptionnels aux administrateurs de leurs réseaux. Il est d'une importance capitale de considérer que l'on **délègue** des **droits d'administration** à son administrateur réseau, et non les pleins pouvoir sur ce que l'on peut considérer comme le savoir stratégique de l'entreprise (le système d'information et les données de l'entreprise). Il est tout aussi important d'inclure dans le cahier des charges de l'administrateur réseau l'**obligation formelle de documenter** tous changements effectués sur le SI ainsi que les différents événements inhérents à son fonctionnement. De plus, il est recommandé d'avoir une liste de spécialistes pouvant intervenir à la demande au cas ou la personne en charge du SI de l'entreprise n'était subitement plus en mesure de le faire. Il devrait être acquis que les personnes exerçant de telles charges (administrateurs réseaux et suppléants) doivent faire l'objet d'une enquête approfondie avant d'avoir accès au SI de l'entreprise.

Le risque informationnel : La société de l'information a bouleversé les rapports de forces classiquement présents dans notre société (du fort au faible), par l'émergence d'un rapport de force asymétrique (du faible au fort). A l'heure actuelle la capacité de nuisance ne se défini plus en terme de puissance d'action et de mise en œuvre, mais plutôt en termes de méthodologies et de compréhension des flux informationnels. La vitesse croissante des échanges (au sein des modèles économiques en vigueurs, entre autres) et la facilité à créer, publier et à faire circuler l'information, de même que l'aspect « de standardisation » de la mise en réseau de données hétérogènes et la montée en puissance des outils de traitement de l'information (data mining[1] et texte mining[2]), permettent, aujourd'hui, de porter atteinte à la plupart des entités économiques et politiques. De plus, la situation économique et une certaine difficulté de compréhension (ou manque de perception) de ces nouveaux modèles d'échanges ont réduit d'autant la marge d'anticipation et la définition même des risques. L'information elle-même s'est dégradée dans sa substance, elle ne bénéficie plus, à l'heure actuelle, des filtres classiquement représentés par la latence due au temps de traitement ou par l'aspect « élitiste » de sa diffusion, car payant auparavant. La gratuité de l'échange à créé le volume par la multiplication des acteurs. Dès lors, le modèle chaotique qui en a résulté ne comportait plus de phase de validation de l'information. Les groupes de presse ont aussi été pris dans la tourmente en épousant un modèle d'économie de marché dont la principale finalité est le rendement. Cette contrainte économique a eu pour conséquence, pour les professionnels de la presse, de diminuer le temps de traitement de l'information, alors que dans le même intervalle il poussait à une professionnalisation des sources ainsi qu'à une réduction de leur nombre. Les répercussions des ces changements n'ont pas encore influé sur les croyances de société qui font que l'on considère encore que « ce qui est écrit est vrai » (par les voies classiques ou électroniques), ou que les images sont des éléments représentatifs de la réalité. Comme on l'a vu récemment dans l'affaire « Borer » ou dans l'actualité des attentats du 11 septembre l'image est un élément qui a gardé sa capacité d'influence, mais perdu sa crédibilité. De même qu'il est facile de créé de l'information et des éléments de preuve.

---

1) Processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données. Un DataMining permet d'analyser les données d'un datawarehouse afin d'extraire des informations originales et des corrélations pertinentes d'un grand volume de données brutes. On parle même de "Découverte de Connaissances dans les Données".

2) Le *text mining* se distingue du *data mining* également par les moyens techniques spécifiques qu'il le faut employer pour traiter les données textuelles et non structurées.

Une définition générale du *text mining* est la suivante : l'extraction d'information à partir des formes ou patrons non manifestes (au sens de *hidden patterns*) dans des grands corpus de textes. Autrement dit, l'objectif est le traitement de grandes quantités d'information qui sont disponibles sous une forme textuelle et non structurée. (Feldman et al., 1998a ; Landau et al.,1998).

Le point N°2 est tiré du document suivant, ce document figure dans les annexe :
TEXT MINING ET INTELLIGENCE ECONOMIQUE : AUJOURD'HUI ET DEMAIN Xavier Polanco Unité de Recherche et Innovation Institut de l'Information Scientifique et Technique Centre National de la Recherche Scientifique

4) L'information aujourd'hui : Internet et le Web, catégorisation et définition

Pour bien comprendre et afin de vulgariser la notion de circulation de l'information on va séparer, de manière très basique, en trois entités distinctes, les principales composantes de la société de l'information :

Internet est le contenant de toutes les informations qui circulent sur le réseau mondial, sa particularité première est ce que l'on pourrait définir comme son mode de langage et d'interprétation : le protocole IP, respectivement TCP/IP (échange par paquet IP). **Internet n'est pas le Web** (www-World Wide Web), mais l'information relative au Web circule à travers Internet. Le principe de fonctionnement de l'Internet est un mode décentralisé et redondant d'échanges d'information par paquets, afin que si l'un des segments du réseau venait à cesser de fonctionner, les segments restants prendraient le relais au niveau de l'acheminement des paquets d'information. Cependant, il faut noter que 80 à 90% des paquets d'information qui transitent sur Internet passent par les Etats-Unis. Les noms de domaine sont la seule ressource qui soit entièrement centralisée. Il y a treize serveurs-racines répartis dans le monde, mais seulement trois qui ne sont pas aux Etats-Unis. Ces serveurs sont de niveau égal, bien que l'un d'entre eux soit appelé "serveur primaire" car c'est sur lui que sont dupliquées en premier lieu les informations qui permettent de faire correspondre les noms de domaine avec les adresses IP. Ces serveurs sont la seule vraie ressource centralisée d'Internet[1]. *Afin de démontrer les possibilités de « causes à effets » de cette « répartition » des « serveurs racine » sur le fonctionnement des entreprises, on peut prendre comme référence l'exemple suivant : entre le 21 et le 22 octobre 2002, les 13 serveurs racines ont subit une attaque simultanée[2] (deny of service – attaque par de multiples messages qui engorgent les serveurs). Bien que le fonctionnement même de l'Internet n'ai pas été mis en danger, on à pu constater une diminution de la capacité à traiter le volume des échanges habituels, ce qui eu pour conséquences de ralentir l'activité économique d'un certain nombre de sociétés. Les auteurs de cette attaque n'ont pas été, jusqu'à ce jour, identifiés, mais de nombreux spécialistes pensent qu'il faut trouver la cause d'une telle action dans l'expression actuelle de la politique étrangère américaine. On peut donc constater que ce type de risque (continuité dans l'accès à l'information) est d'ordre nouveau et dépasse complètement le « périmètre » habituel de surveillance de l'entreprise.*

Le protocole (TCP/IP) ne peut pas être considéré comme fiable dans sa version actuelle (IPv4), celle-ci permet, entre autres, les usurpations d'identités et d'origines géographiques (pour les adresses email ou les sites Web par exemple). Il faut prendre en compte que pour changer cette situation, une version améliorée du protocole IP (Ipv6) va être prochainement implémentée. L'organisme de standardisation du protocole IP est l'IETF.

Le Web gère les échanges conventionnels d'informations auxquels nous sommes habitués à accéder par le biais d'un navigateur (Browser). La particularité du Web est sa fonction « hypertexte » qui offre la possibilité de lier un document, un mot ou un élément à un autre, quel que soit son emplacement géographique. Ses principaux modes de transports de l'information sont le protocole « http » et le langage de description de document « html » (ces éléments et les risques encourus au niveau de l'information sont traités spécifiquement dans la partie « aspect multicouche de l'information »). Les principales failles de sécurités présentes au niveau du Web sont dues à un manque de suivi des standards en place par les concepteurs de logiciels, ainsi qu'à la nécessité, et le manque de fiabilité, des mises à jour des SI et programmes associés. Il en résulte qu'il est possible d'exécuter des scripts malveillants par le biais des navigateurs Web.

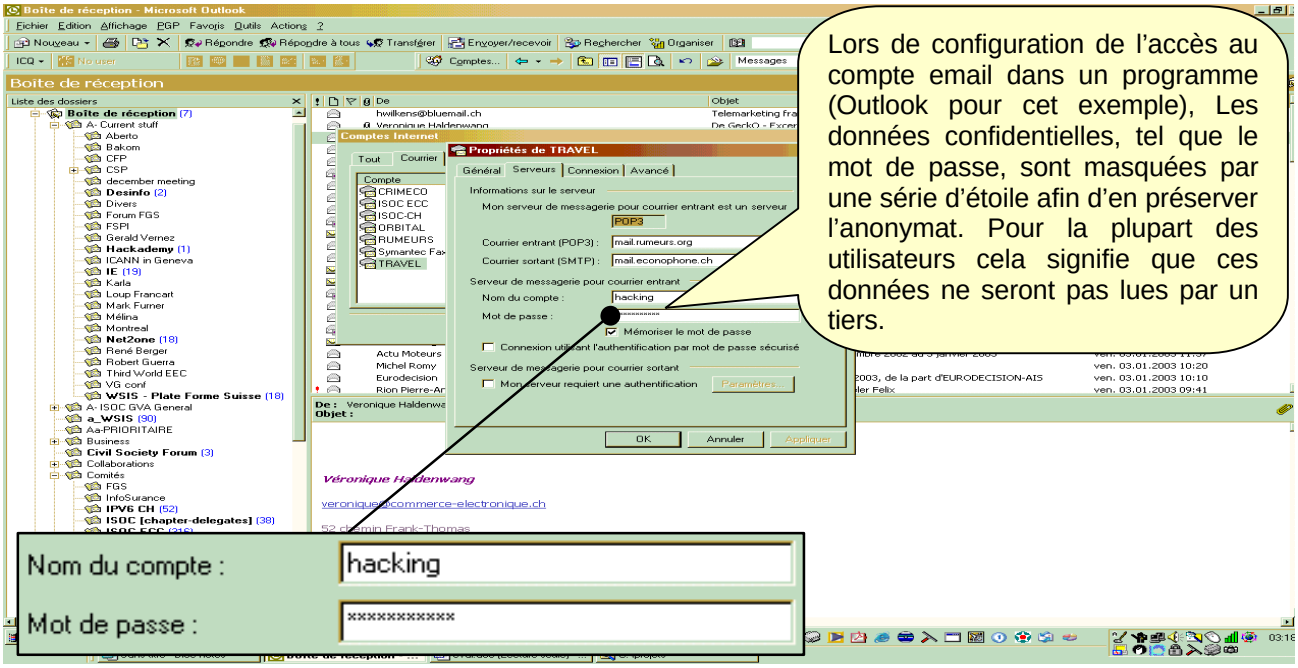
Intégrés dans l'environnement direct du Web on peut noter une série de protocoles : l'email (SMTP. POP, IMAP), le téléchargement de fichiers (FTP), les groupes de news (USENET) les forums de discussions et messagerie instantanées (IRC / ICQ / Messenger / Chat), les systèmes de messages BBS (Bulletin Board System), ou encore la connection à une autre machine (TELNET).

1) Root servers, vois annexe : « Répartition géo-stratégique de l'Internet »
2) Journal du Net : http://solutions.journaldunet.com/0210/021024_rootserver.shtml

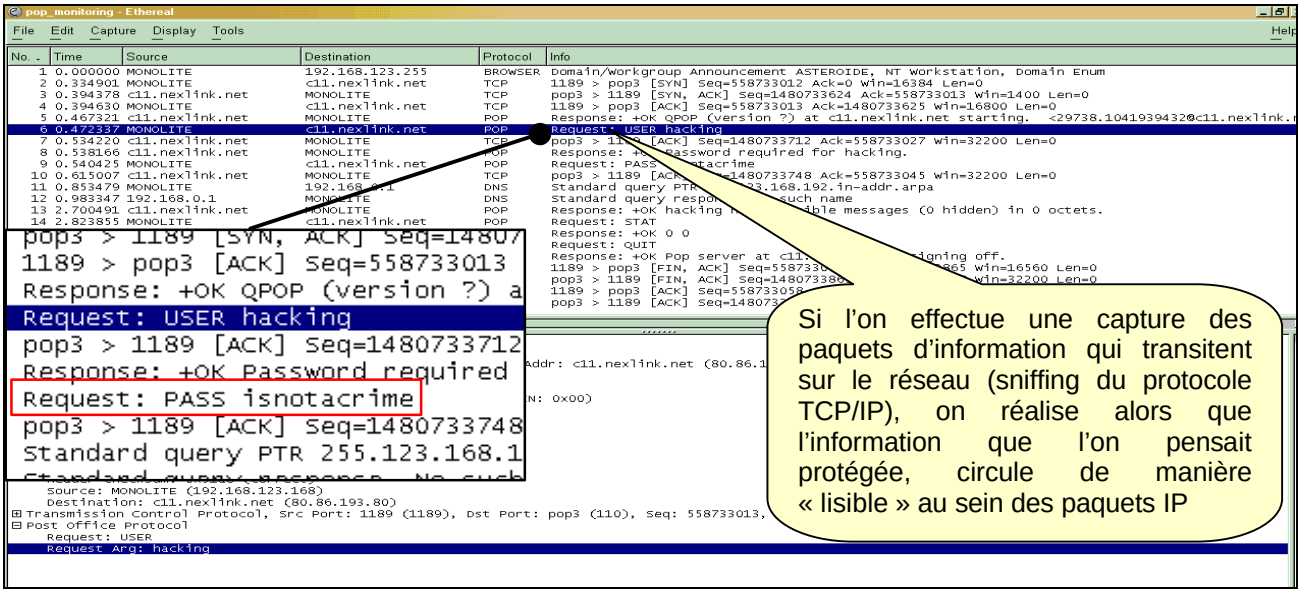
5) L'aspect multicouche de l'information

L'aspect multicouche de l'information : les paquets IP

Une des problématiques au niveau d'Internet et du Web concerne les possibilités multiples de fraudes au niveau de la diffusion de fausses informations, d'usurpation d'identité ou de la publication de faux sites Web. Une des facilités de mise en œuvre de ce type de fraude est la méconnaissance de certains aspects techniques par les utilisateurs. De manière générale l'adage « je crois ce que je vois » fonctionnait très bien dans notre société classique, mais malheureusement avec l'avènement du Net les choses ont changé. L'exemple ci-dessous a pour but de démontrer la différence entre le contenu auquel on accède et l'information qui est réellement transportée dans les paquets IP.



Pour capturer ces informations il est nécessaire d'accéder au réseau de l'entreprise, mais dans le cas d'un système sans fils types Wireless lan (802.11), cette opération sera aisée dans des conférences ou des lieux publics ou ce type de système est généralement configuré sans restriction d'accès.



5.1) L'aspect multicouche de l'information: Le Web (et quelques protocoles associés)

En ce qui concerne le Web, la fraude ou la tromperie se fait généralement au niveau du langage « html » ne demandant pas un haut niveau de connaissances techniques de la part du fraudeur pour leur réalisation. Mais pour les même raisons que celles citées précédemment, elles ont toutes les chances d'aboutir avec des utilisateurs « non-éduqués ».

Le protocole utilisé pour l'accès à un document révèle la nature de ce dernier. Par exemple, sur un serveur *FTP (file transfer protocole)*, les ressources proposées sont destinées au téléchargement. Ces protocoles ne nous apportent pas d'information concernant le contenu du document, mais ils donnent une information sur la **nature**, la **forme**, le **format électronique** du document disponible en ligne.

Voici les principaux protocoles auxquels on peut être confronté sur le WWW ainsi qu'à l'intérieur des pages HTML, par l'intermédiaire des liens hypertextes :

http://	Hypertext Transfer Protocol : protocole de communication utilisé pour les échanges de données entre les clients et les serveurs « www »
ftp://	File Transfer Protocol : protocole de transfert de fichiers entre deux machines sur Internet
gopher://	Protocole aujourd'hui supplanté par le « http:// », système d'information distribué, l'accès à l'information est structuré selon un réseau de menus multi-niveaux
telnet://	Protocole d'application définissant l'émulation d'un terminal sur Internet
mailto:	Protocole d'accès d'un e-mail

HTML[14] (HyperText Markup Language) n'est pas un langage de programmation ! Ce n'est qu'un langage de description de documents, il est utilisé pour écrire les pages standards du Web).

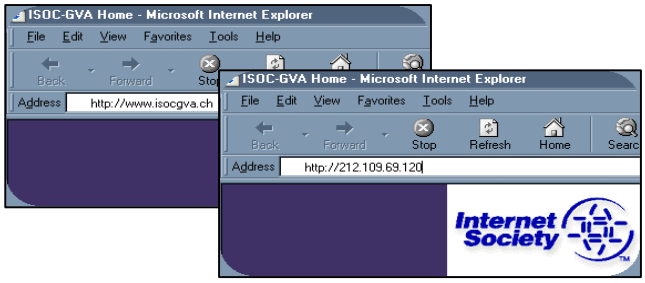
En d'autres termes, HTML est un ensemble (réduit) de balises (ou styles ou "tags") utilisés pour définir les différents composants d'un document.

L'accès au document, et la définition de son emplacement se fait grâce à un **URL** (*Uniform Ressource Locator*). Le nom du document est précédé par son chemin d'accès, le point de départ de celui-ci étant représenté par un nom de domaine [ex: www.switch.ch] ou le numéro IP d'un domaine [ex: 192.247.93.18] en ce qui concerne un document online:

Le plus souvent un **URL** sera de la forme : http://nom_de_domaine/nom_de_document

5.2) L'aspect multicouche de l'information, l'adresse IP

L'adresse IP sous sa forme chiffrée ou le nom de domaine Internet (domain names system, DNS), qui servent à identifier un site Internet, peuvent être utilisés indifféremment pour accéder au site en question, on les considère comme des adresses. C'est la raison pour laquelle le terme « adresse Internet / URL » est largement utilisé pour désigner ces deux notions.



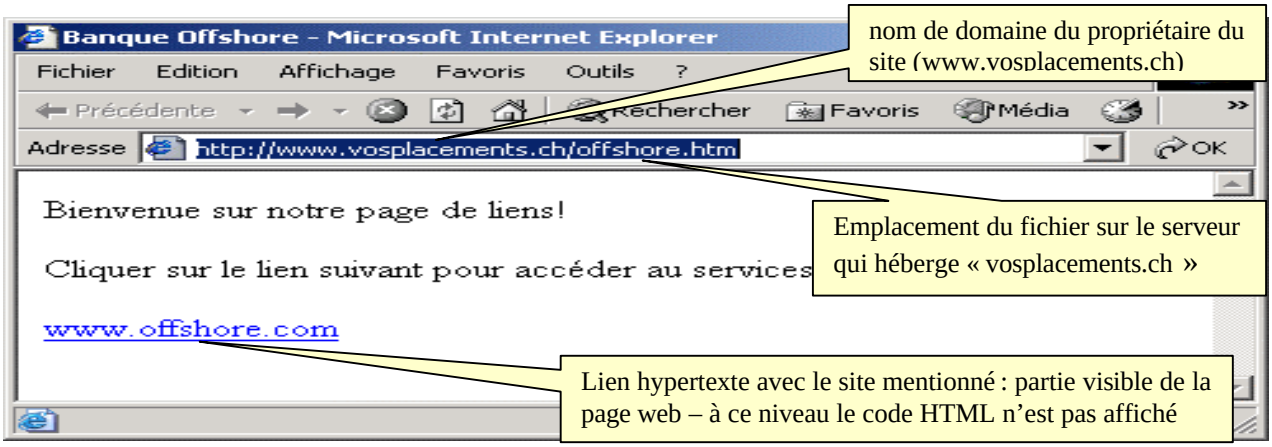
À titre d'exemple, l'adresse IP **132.203.250.87** peut se lire ainsi : l'ordinateur 87 situé sur le 250e réseau du réseau 203 qui, lui, se trouve dans le **réseau global 132** (ou par exemple : l'ordinateur qui se trouve rue de la Gare 10 [87], à Genève [250], une ville qui se trouve en Suisse [203], un pays qui se situe **en Europe [132]**). Ce type d'adressage constitue la base du protocole de communication TCP-IP. Il permet aussi la localisation de la machine qui héberge le mon de domaine et généralement le contenu du site Internet. Normalement des classes d'adresses sont attribuées par pays, mais il est extrêmement difficile de localiser géographiquement, de manière précise, une adresse IP (ce sujet sera abordé dans le chapitre sur la recherche d'information).

5.3) Exemple de redirection d'un site Web dans le but de commettre une fraude :

CNN a récemment été victime de ce type de fraude : *« Les fausses pages générées ressemblaient à s'y se méprendre à celle du site de CNN. Le logo, les liens vers les rubriques et les derniers articles, tout y était. Y compris l'adresse de la fausse page, qui commençait par http://www.cnn.com, suivi d'une arobase et d'une adresse IP. Le néophyte n'y voyait que du feu. Et beaucoup ont cru aux fausses infos. »*[15]

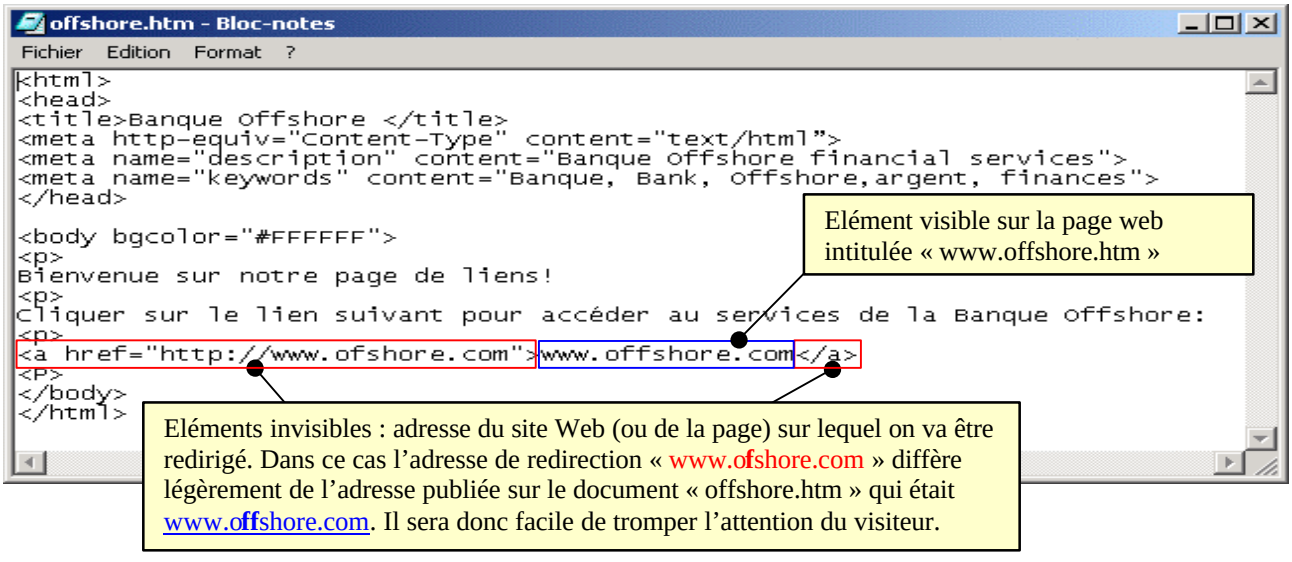
Pour illustrer cet exemple, je vais publier le document: [offshore.htm] sur le Net et le lier ensuite avec la page d'un site Web [www.vosplacements.ch]:

Voici la page telle qu'elle apparaîtrait dans un navigateur :



Voici la page telle qu'elle est en réalité :

*Démarche pour afficher le code source de la page **« offshore.htm »** : Dans Internet Explorer : sélectionner le menu « **Affichage** », ensuite sélectionner dans l'arborescence : « **Source** ») :*



15) Source: TF1 : http://www.tf1.fr/news/multimedia/0%2C%2C986128%2C00.htm

6) La recherche d'information sur le WEB

6.1) Les principales sources d'information :

Sources blanches : basée essentiellement sur des sources ouvertes et libres d'accès, cette catégorie à connue un grand essor avec le développement de l'Internet. On peut trouver un grand nombre d'information sur les sociétés, les individus, le savoir académique et le contenu des médias électronique. Le nombre de source et l'absence de contrôle de celle-ci implicite une utilisation prudente des résultats des recherches effectuées.

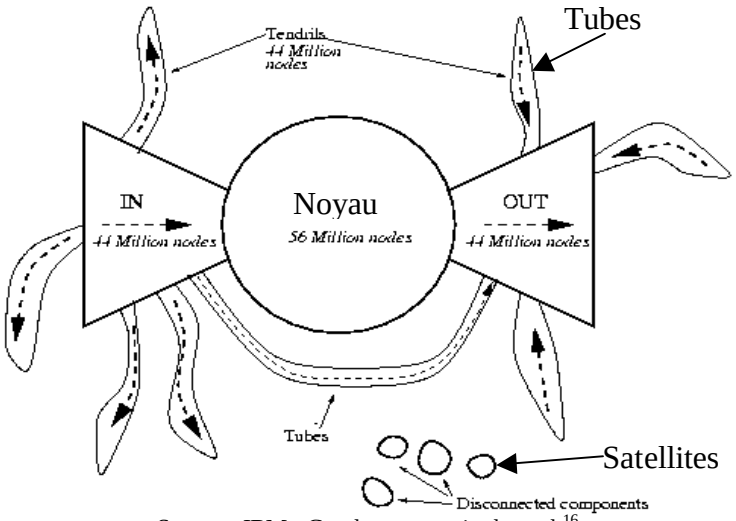
Sources professionnelles : c'est une sous catégorie des sources ouvertes, elles ne sont accessibles que contre paiement (à l'information ou à la quantité), cependant au contraire des sources libres d'accès, elles font, dans la majorité des cas, l'objet d'un contrôle au niveau de la qualité et de la crédibilité de l'information stockée ou traitée. Le principal problème des sources professionnelles c'est qu'elles ne référencent que les publications officielles. C'est pour cette raison qu'il est nécessaire de s'intéresser à l'ensemble des sources potentiellement disponibles.

Sources grises : C'est le coté non formalisé et non explicite de l'information. Cela pourrait être ce que l'on va recueillir lors d'un séminaire ou d'une conversation. Ou cela peut aussi être représenté par des sources d'information inaccessibles aux différents types de moteurs de recherches ou encore par des sociétés de conseil qui en plus de procéder à la recherche, font passé les résultats par un réseau d'experts ou de consultants qui valorisent l'information en fonction de leurs propres connaissances.

Sources noires : Ce type d'information est, dans la majorité des cas, obtenue de manière illégale ou d'accès fermé. Dans cette catégorie on range l'information collectée de manière indue, telle que l'achat de renseignements consternant une entreprise (espionnage économique).

6.2) Cartographie de l'information disponible sur le Web :

Les moteurs de recherches conventionnels ne trouvent que 20 à 25% du contenu présent sur le Web. Le graphique ci-dessous montre que l'information n'est pas répartie ou accessible de manière homogène : le « Noyau » représente les sites Web et bases de données interconnectées (par des liens hypertextes). Certains sites sont connectés à ce « Noyau » depuis l'extérieur (IN), d'autres sites sont connectés du « Noyau » vers l'extérieur (OUT). Des sites sont connectés aux éléments extérieurs sans être en liaison avec le « Noyau » (Tubes). Et finalement, certains sites ne sont pas du tout connectés à d'autres.



Source : IBM : Graph structure in the web[16]

16) IBM : Graph structure in the web : http://www.almaden.ibm.com/cs/k53/www9.final/

Les 75 à 80% de l'information restante (celle que les moteurs de recherches n'arrivent pas à indexer) représentent ce que l'on appelle « le Web invisible ». Celui-ci est principalement composé de bases de données universitaires, de sites « satellites », ainsi que de site « dynamiques » (dont le contenu est généré à la demande). Pour les recherches d'information dans ces deux parties (visible et invisible) la méthodologie reste la même, seuls les outils varient quelque peu.

La durée limitée dans le temps de la présence de certaines informations sur le Web peu rendre l'exercice fastidieux. Pour cette raison, il est souvent utile de conserver une copie des informations utiles au moyen d'outils appropriés (Acrobat Exchange, pour fabriquer des documents PDF, ou un aspirateur de site pour conserver une copie du site désiré)

Estimation de la taille du Web (mars 2002)
- 13 milliards de documents
- 7,5 millions de nouvelles pages par jour
- 50 à 75 terabytes d 'information
- 600 milliards de pages dans l'ensemble des sites intranet

Répartition des Noms de domaine
".com" : 54,68% - ".org" : 4,35%
".net" : 7,82% - ".gov" : 1,15%
- ".edu" : 6,69%

Langues des documents
- 56,6% des pages sont en Anglais
- 2,4% en français,
- 0,5% en allemand.

Source : Recherche d'information et veille sur Internet (www.enpc.fr/enseignements/Legait/projet/victor/chercher/sources.html)

6.3) Les types d'outils et leurs fonctions[17] :

Il existe des milliers d'outils de recherches, dans les exemples qui seront donnés on va se concentrer sur un choix non-exhaustif des moteurs et outils de recherches les plus performants dans le contexte de ce travail. Il faut néanmoins garder à l'esprit que le choix du moteur de recherches fait partie de la « stratégie de recherche ». Pour arriver à des résultats probants, il faut en premier lieu faire des « recherches sur les outils de recherches », celles-ci peuvent s'effectuées par l'intermédiaire des moteurs conventionnels au moyen d'une requête adéquate : par exemple : [*moteur de recherches des adresses email* / *email address search engine*]. Mais le meilleur moyen reste l'utilisation des annuaires ou portails thématiques qui contiennent déjà toute l'arborescence des outils de recherches :

Les principaux annuaire thématiques sur les moteurs de recherches (MR)
http://outils.abondance.com
http://c.asselin.free.fr/french/moteurs.htm
www.adbs.fr/site/repertoires/sites/lardy/outils.htm
www.liensutiles.org/rechspec.htm
www.searchenginewatch.com/links
www.searchtools.com
Les principaux annuaire thématiques concernant le Web invisible
http://c.asselin.free.fr/french/webinvisible2.htm
www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html
www.invisible-web.net

6.4) Les différents outils de recherches de l'information:

Les annuaires : leur particularité est de regrouper thématiquement des sites Web dans des répertoires addoc. Dans la majorité des cas, le catalogage et l'indexages des sites ce fait pas des personnes *(au contraire de l'indexation des sites fait par des programmes automatiques pour les MR par mots-clés).* Les ***avantages*** de ce type d'outil sont une bonne pertinence des résultats obtenus ainsi qu'une meilleure maîtrise de l'environnement de recherche (ont accède à un nombre de résultat limités, mais très ciblés). Les ***inconvénients*** de la recherche par navigation sont des champs limités par le contenu des répertoires auxquels on accède, ce qui nécessite de savoir très précisément ce que l'on cherche. [*par exemple : http://dmoz.org*]

Les métas-moteurs : leur particularité est d'effectuer les dites recherches au travers d'autres MR, les requêtes sont basées sur des mots-clés (une requête sera transmise à plusieurs moteurs de recherches en même temps). L'***avantage*** de ce type de recherche est de couvrir un champs très large et de pouvoir utiliser un nombre important d'outils lors de la même requête. L'***inconvénient*** de ces métas-moteurs se situe au niveau des résultats des requêtes : il y non seulement une redondance d'information, mais en plus la pertinence des résultats obtenus laisse à désirer. On les utilisera pour des recherches généralistes. [*par exemple : www.37.com*]

Les moteurs de recherches par mots-clés : ils sont composés d'un ensemble de bases de données crées de manière automatique par des programmes spécialisés (crawlers ou robots, qui parcourent le Web à la recherche de nouveaux sites). La fréquence de mise à jour de l'index dépend de la taille de la base de données et de la politique du MR. Ces MR utilisés tels quels, ont les mêmes avantages et inconvénients que les métas moteurs. Leur ***avantage***, c'est les fonctionnalités avancées que l'on peut exploiter dans les requêtes, chaque MR possède une page spécifique pour ces fonctionnalités (ainsi que pour les opérateurs « type Booléens » de recherches disponibles tels que « and », « or », « near » etc. voir liste au point 6.8), qui bien utilisées permettent d'obtenir des résultats très pertinents. L'***inconvénient***, dans leurs fonctions de bases ils ne sont pas très performants (si l'on n'utilise pas les opérateurs de recherches – décrits sous rubrique d'aide du MR). [*par exemple : www.google.com*]

Les moteurs de recherches spécialisés : qu'ils se présentent sous formes d'annuaires ou de recherches par mots-clés, ils ne couvrent qu'un seul domaine (par exemple pour les recherches de sociétés, d'emails, de documents PDF, d'images, etc.). Ils permettent une meilleure qualité dans les résultats des requêtes. [*par exemple : www.societes.com, pour les entreprises, ou encore www.phonenumbers.net pour les numéros de téléphone dans le monde*]

Les moteurs de recherches de cartographie de l'information : ces MR qui fonctionnent par mots-clés, n'affichent pas une liste de résultats lorsque l'on fait une requête, mais ils affichent une carte de l'information obtenue, le schéma auquel l'on va accéder est une représentation graphique de la requête, de son résultat et de l'environnement de l'information en relation avec la dite requête. Ces outils de nouvelle génération apportent une convivialité dans la recherche ainsi qu'une représentation visuelle qui permet d'améliorer son approche de l'information (d'autres outils d'interprétation et de visualisation de l'information seront présentés dans la partie 6.5). [*par exemple : www.kartoo.com ou http://maps.map.net (qui permet de cartographier un annuaire tel que dmoz.org)*]

Les outils humains de recherches : ces MR sont basés sur une aide "humaine". Les questions parviennent à une équipe de spécialistes des outils de recherches qui formulent, à la demande, les requêtes de recherches les plus adéquates. Ce genre d'initiative permet au non-initié de trouver des informations plus facilement, et elle permet aussi au spécialiste d'élargir sa palette d'outils (et de compétences). Certains de ces sites sont gratuits [*par exemple : www.webhelp.fr*]

Les réseaux experts : ces réseaux ne se composent pas à proprement parlé d'un moteur de recherches, ils se servent plutôt du Web comme d'une plate-forme d'accueil pour réceptionner les requêtes de leurs clients. Ils ont un réseau de spécialistes de la recherche et ils croisent les résultats

avec des réseaux d'experts (consultants spécialisés dans des secteurs d'activités distincts) pour créer une information à valeur ajoutée. Ce type de réseau, que l'on peut assimiler à l'intelligence économique au niveau du traitement de l'information, permet d'accéder à une catégorie d'information « grise » et « professionnelle ». Des sociétés de services comme SVP Conseil sont abonnées à une multitude de bases de données professionnelles telles que, par exemple, Factiva, Lexis-Nexis, Dialog, etc… Elles vendent aussi la possibilité d'accéder à celles-ci par mois ou par zone nationale. Le type d'information que l'on peut obtenir par le biais de ce type de réseau d'experts représente une haute valeur ajoutée par rapport à l'utilisation simple des sources ouvertes. [*par exemple : www.svp.fr, www.egideria.fr, www.world-check.com, www.insideco.net, www.krollworldwide.com*]

On fera une petite parenthèse pour les réseaux experts et bases de données telles que *www.world-check.com et www.insideco.net,* ces bases servent surtout à des aspects de Due Diligence sur des personnes (politiquement exposées, entre autres) ou sur des sociétés considérées comme douteuses. Il faut tenir compte que ce type d'information ne peut pas être considéré comme une information de première main (le contrôle des sources est extrêmement difficile), de plus, suivant le cadre législatif du pays ou l'on désire exploiter les données fournies par ces organismes, on peut se trouver confronté à un risque juridique (cadre légal sur la protection des données personnelles) lorsque l'on voudra utiliser ces données. Il est recommander, pour des cas sensibles, de faire appel à du conseil spécialisé (rapport de renseignement par des professionnels qualifiés), à ce titre, il ne faudra pas négliger de faire appel aux structures administratives existantes (police, renseignement, ambassades..).

6.5) Autres outils de recherches et traitement de l'information :

Source et url de référence : www.enpc.fr/enseignements/Legait/projet/victor/chercher/Outils.html

les outils de surveillance " tracking " et d'alerte "cyberalert " : Ils sont des fonctions de recherche, de présentation et de distribution de l'information. Il y a deux options pour faire la surveillance: par abonnement gratuit à un site de surveillance ou bien par un logiciel de surveillance

les outils : Agents intelligents : Ils remplissent plus ou moins en profondeur les sept fonctions : Rechercher, Indexer, Filtrer, Sauvegarder, Présenter, Distribuer, Aider à la décision. Pour une surveillance optimale sur Internet

les outils d'aspiration " mirroring ": Ils remplissent les fonctions de sauvegarde (recopie), d'indexation et de filtrage pour certain d'entre eux. Ils dupliquent tout ou partie d'un site en recopiant les pages, les répertoires et l'arborescence du site sur le poste informatique local

les outils de gestion intelligente " Information mining ": Ils réalisent les fonctions : Rechercher, Indexer, Filtrer, Sauvegarder, Présenter, Distribuer, Aider à la décision. Ils sont centrés sur la gestion intelligente de l'information même s'ils intègrent de plus en plus des fonctions de recherche sur le web

les outils de d'analyse et de représentation de l'information: Ils ont les fonctions : Indexer, Filtrer, Sauvegarder, Présenter, aide à la lecture d'un ensemble de documents sur le web sous forme de représentation graphique, qui fournissent une meilleure compréhension rapide de grand volume d'information. [*voir aussi : http://c.asselin.free.fr/french/carto.htm*]

NdL : Pour accéder à la liste des différentes catégories «d'outils de recherches et traitement de l'information » disponibles veuillez consulter l'url de référence (source). Certains de ces outils peuvent ne plus êtres accessibles en raison de considérations d'ordre économiques.

6.6) La définition des zones et périmètres de recherches :

Le mode d'emploi des fonctionnalités qui figurent ci-dessous figure généralement sous la rubrique « recherches avancées » des MR, il faut aussi noter que les MR comportent une rubrique d' aide

qui décrit leur mode de fonctionnement.

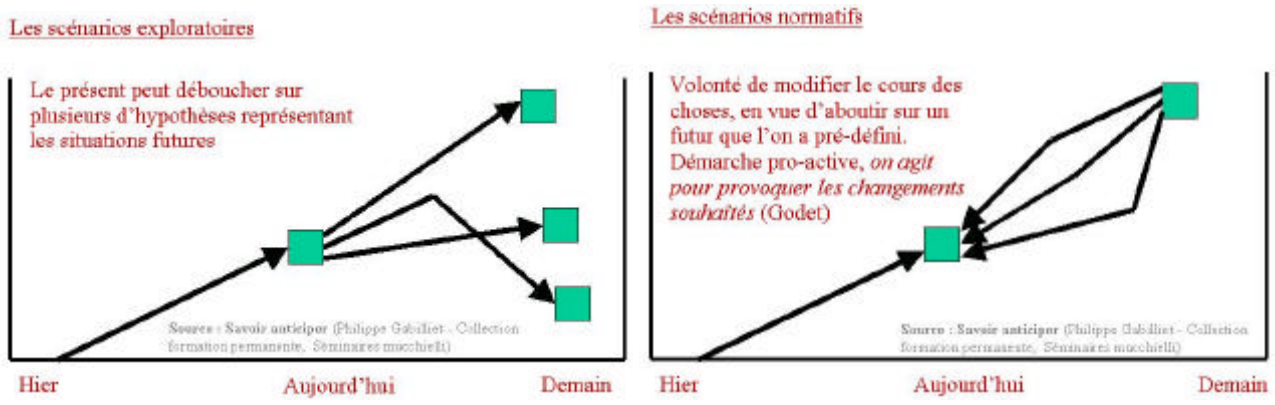
On peut effectuer des recherches à l'intérieur d'un site complet ou uniquement dans une page Web. Il est possible de ne s'intéresser qu'aux liens visibles sur la page Web, ou à la l'information qu'ils contiennent dans la description des liens qui figure dans les balises html. On peut effectuer une recherche par rapport au nom de domaine (par pays ou par un générique, tel que .com) ou par rapport au nom d'un site. On peut aussi chercher par le type et format de l'information désirée (image, vidéo, document world, PDF, Excel, etc). Certains MR, tel que Google [*www.google.com*], possèdent une fonction « cache» qui permet d'accéder à une version enregistrée du document, même s'il n'est plus disponible sur le serveur d'origine. Cette fonction est aussi disponible avec « the Wayback Machine [*www.archive.org*]» qui intègre, depuis 1996, plus de 10 milliards de pages d'archives. Il est aussi possible de chercher au travers des « en-têtes Meta » par les « keywords » ou les « description ». On peut aussi faire des recherches centrées sur des éléments spécifiques tels que les carnets d'adresses présents sur les page de liens des site Web [exemple de formulation de requête : *url:bookmark*].

6.7) Principes de bases pour établir une recherche :

Il faut être conscient que le fait de rechercher une information donne déjà une information (vers l'extérieur) sur nos centres d'intérêts. Ce problème sera abordé dans la troisième partie : les méthodes de protection lors du traitement d'informations sensibles.

Dans un premier temps on va aborder la recherche sous l'angle logique (il faut s'ouvrir l'esprit, et non s'arrêter à la perception technique des outils) et établir une stratégie de recherche : nature de l'information recherchée, le type concerné, objectif de la recherche, type de résultat souhaité, temps à disposition. On va commencer à délimiter l'environnement de sa recherche : que l'on cherche exactement (une personne, un document, un format spécifique, un email) ? Est-ce que le type d'information recherché est ancien ou récent, plutôt d'ordre académique, professionnel ou privé, est-il en relation avec les sources blanches ou grises ? Quels sont les éléments utiles en relation avec la requête (est-ce l'on dispose d'autres éléments « d'information indirecte » qui permettraient d'effectuer une recherche en parallèle afin de trouver des relations avec ma recherche principale, et déterminer les ressources les plus pertinentes) ? Quel est l'éventail des outils dont on peut disposer (selon la nature de la requête, quel est l'outil le plus approprié : un méta-moteur, un annuaire, un newsgroup, une base de donnée professionnelle) ? Combien de temps est-ce que je peux consacrer à ma recherche, est-ce que les coûts sont en rapport avec les résultats ?

<u>Exemple de deux approches logiques :</u>



Il faut ensuite interpréter le mode de fonctionnement de l'outil choisit comprendre sa manière d'effectuer une recherche. On va donner du sens à sa requête (expression) et utiliser un maximum de mots en relation avec ladite requête, de cette manière le nombre résultats à filtrer sera réduit au minimum : définir les éléments qui vont composer la requête en partant du général au particulier.

Ceci va améliorer la pertinence des résultats et de diminuer au maximum « le bruit » . Il est des fois utile d'utiliser plusieurs moteurs en parallèle (suivant leurs spécificités et efficacités). A l'obtention des résultats des recherches il sera nécessaire de faire un travail de sélection et de hiérarchisation de l'information ainsi obtenue. Par la suite il faudra de nouveau structurer sa requête (title, host, url, …) et redéfinir ses besoins en information par rapport au recherches précédentes (est-ce que des éléments utiles peuvent venir se greffer dans les prochaines requêtes, est-ce que ma manière de présenter ma requêtes est adéquate ou est-ce que je dois changer l'orientation de celle-ci au vu des résultats obtenus ?).

6.8) Utilisation des fonctionnalités avancées des moteurs de recherches : les principaux « Opérateurs Booléens » :

ADJ – à côté de

Utilisez ADJ pour retrouver deux mots côte à côte dans l'ordre.

AND - et

Utilisez AND pour retrouver des notices qui contiennent deux termes.

IN – dans

Utilisez IN pour rechercher un mot champ particulier dans un champ spécifique.

NEAR – à proximité de

Utilisez NEAR pour retrouver les enregistrements qui contiennent les deux termes dans la même phrase.

NOT – pas de

Utilisez NOT pour rechercher des enregistrements qui contiennent l'un des termes seulement.

OR - ou

Utilisez OR pour rechercher des enregistrements qui contiennent l'un, l'autre, ou les deux termes.

WITH - avec

Utilisez WITH pour rechercher des enregistrements qui contiennent les deux termes dans le même champ.

WILDCARD – troncature à l'intérieur du mot

Sert à remplacer une lettre ou représente l'absence d'une lettre (quand on a un doute sur l'épellation). Ex. m?cdonald retrouve à la fois mcdonald et macdonald; p??re retrouve père et paire.

Troncature

Vous pouvez utiliser le symbole de troncature (*) pour remplacer un caractère ou une chaîne de caractères. Ex. cat retrouve les mots catégorie, catatonie, cats, etc.*

7) L'analyse et la crédibilité de l'information (inclus traçabilité et identification)

La liberté et la facilité de publication de documents électroniques sur le Web sont significatives de la difficulté représentée par l'identification de la source des informations auxquelles on peut accéder. C'est pour cette raison, qu'en plus des **démarches « logiques »** de contrôle (*date de l'information, date de mise à jour, notoriété et crédibilité de la source, possibilité de contacter son émetteur, qualité de l'url*), il faudra entamer des **démarches « techniques »** liées à l'identification de la provenance de l'information (géographique, technique/réseaux, source et traçabilité d'un email ou d'un site web), à l'étude de la structure de l'information (code html d'une page Web, d'un email html), à l'identification de l'ayant droit économique (détenteur d'un site Web), ou encore pour déterminer le prestataire de service qui héberge le nom de domaine, ainsi que le contenu d'un site Web. Ces démarches font appel à une complémentarité des connaissances abordées précédemment (aspect multicouches de l'information, notion d'adressage IP, url, code html, méta tag).

7.1) Crédibilité de l'information, présentation de deux cas école

a) Le cas Emulex *: cette société active dans le domaine de la fibre optique a été victime en août 2000 de la propagation d'une série de fausses informations la concernant. Alors que le marché boursier du Nasdaq ouvrait à peine ses portes, l'action de la société Emulex s'effondrait brutalement,* ***projetant le cours de l'action de 103 à 45 dollars en l'espace de quelques minutes,*** *ce qui provoqua une perte estimée au niveau du capital action à 200 millions de dollars ! A l'origine de cette baisse notable se trouve un communiqué,* ***repris par Bloomberg et*** *Downjones (deux des trois plus importants brokers d'informations financières au monde). Ce communiqué affirmait le plus sérieusement du monde que les résultats escomptés ne seraient pas à l'ordre du jour, que Paul Folino, patron de la firme, démissionnait et même que les autorités boursières avaient lancé une série d'enquêtes sur les comptes de la société. Ces information étaient totalement erronées, cette fraude avait été organisée par un ancien employé de Downjones, lequel a nourri les deux services susmentionnés avec des fausses données. Seul Reuter, grâce à des méthode de validations des sources en amont (seul la réception de la validation de l'information par une source tiers permet à la « news » d'aller plus loin) à réussi à « filtrer » ces fausses informations.*

b) ***Le cas CNN*** *: (exemple abordé à la page 8 pour ce qui est de la redirection d'un site Web vers une copie illégale de celui-ci). Cette affaire regroupe les aspects techniques liés à la méconnaissance des utilisateurs, ainsi que les aspects de crédibilité du au rapport de confiance induit par le fait que l'on pensait se trouver sur le site de CNN, et donc que l'information ne nécessitait pas de contrôle supplémentaire. Le type d'url utilisé permet de rediriger l'adresse d'origine du site Internet cible sur le site contenant les fausses informations. La charte graphique étant reproduite à l'identique, la plupart des gens se sont faire prendre à défaut. Les professionnels de la presse eux-même n'ont pas réussi à détecter la fraude, un nombre considérable de nouvelles ont été reprise in extenso sur des sites – officiels – tiers. Crédibilisant des informations déjà fausses à l'origine. Donc les personnes qui ont accédé aux sites web qui avaient repris l'information à leur compte n'avaient que des signaux très faibles de remise en cause de l'information*

Exemple de l'url utilisée dans le cas CNN :

http://cnn.com:443@212.190.116.226/news.php?y2JEHUDv

Dans ce cas, plusieurs journaux, ainsi que l'université de Notre Dame ont été dans l'obligation de publier des démentis

le « **:443@** » permet une redirection depuis n'importe quel site Web. Il suffit donc d'imiter la charte graphique du site cible, et ensuite d'obtenir l'adresse IP de son propre domaine [*212.190.116.226]*, (voir page 7, « l'aspect multicouche de l'information »), l'inclure à la suite de l'arobase avec le chemin jusqu'au document désiré [*/news.php?y2JEHUDv*].

Ces deux cas illustrent aussi très bien le concept de « rapport de force asymétrique » Pour chacun d'entre eux, une seule personne a œuvré et à réussi à compromettre la vie d'une entreprise cotée en bourse pour le premier et a déstabiliser une chaîne d'information telle que CNN. La différence

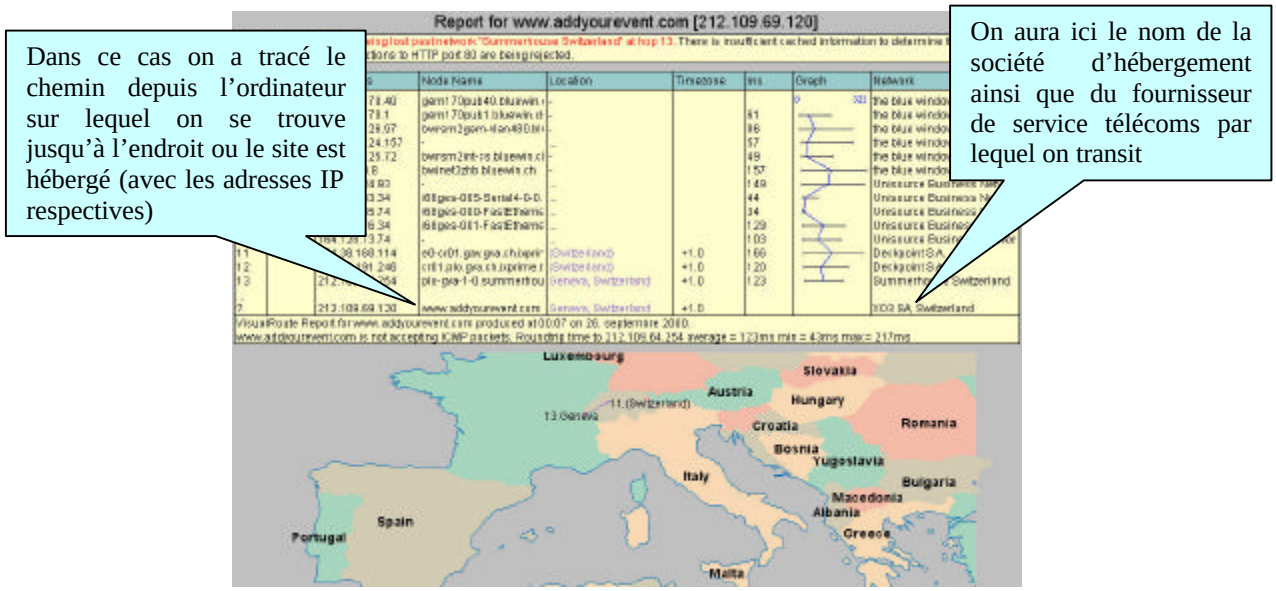
entre les moyens engagés et la puissances des entreprises attaquées est la représentation de ce rapport asymétrique

7.2) Crédibilité de l'information : les démarches de validation « logiques »

- **date de l'information** *: quand est-ce que l'information a été publiée, est-ce que la date de publication correspond aux autres dates présentes dans le site ? – quand disponible…* [Information quantitative]
- **Date de mise à jour** *: est-ce que le site fait l'œuvre d'une politique de mise à jour de l'information, est-ce que celle-ci semble homogène sur l'ensemble du site ?* [Information quantitative]
- **Notoriété et crédibilité de la source** *: est-ce que le site est connu, de quelle notoriété bénéficie-t-il, quels sont les résultats que j'obtiens quand je le soumets à des MR (ce type de contrôle est possible avec des outils spécialisés tel que : www.linkpopularity.com)* [Information quantitative]
- **possibilité de contacter son émetteur** *: est-ce que l'information à laquelle j'accède est signée ou légendée, est-ce que les coordonnées de l'auteur figurent sur le site, est-ce que les informations de contacts sont composées d'email, d'adresse postale, d'un numéro téléphone ou de télécopie ?Quels sont les résultat lors de recherches d'après les informations de contact figurant sur le site (prenez les nom (personnes ou sociétés) et avec un moteur de recherche ou un meta-moteur, essayer de trouver des " traces " de celles-ci)* [Information qualitative]
- **structure de l'url***: (selon l'exemple de CNN)est-ce que le nom de domaine est la propriété de l'éditeur, ou est-ce que l'on a affaire à un site qui offre des espaces d'expression gratuit, est-ce que la dénomination des liens correspond bien à la structure des url, est-ce que l'extension du nom de domaine est connue et en relation avec l'emplacement géographique supposé de l'information, ou est-ce que l'on a affaire un nom exotique (par exemple un www.nom.fr.st pour un article en français) ?* [Information qualitative]
- **référencement de l'information contenue dans le document** *: est-ce que toutes les informations en relation avec le document auquel l'on accède sont bien documentées, les sources citées et vérifiables (lien avec la source) est-ce que les personnes citées figurent sur le Web (prenez le nom et prénom et avec un moteur de recherche ou un meta-moteur, trouver les " traces " de cette personne)?* [Information qualitative]
- **possibilité de croiser l'information avec d'autres sources d'information** *: est-ce que l'on peut trouver trace de l'information à laquelle on accède dans d'autres sources, le message est-il homogène ?* [Information quantitative]

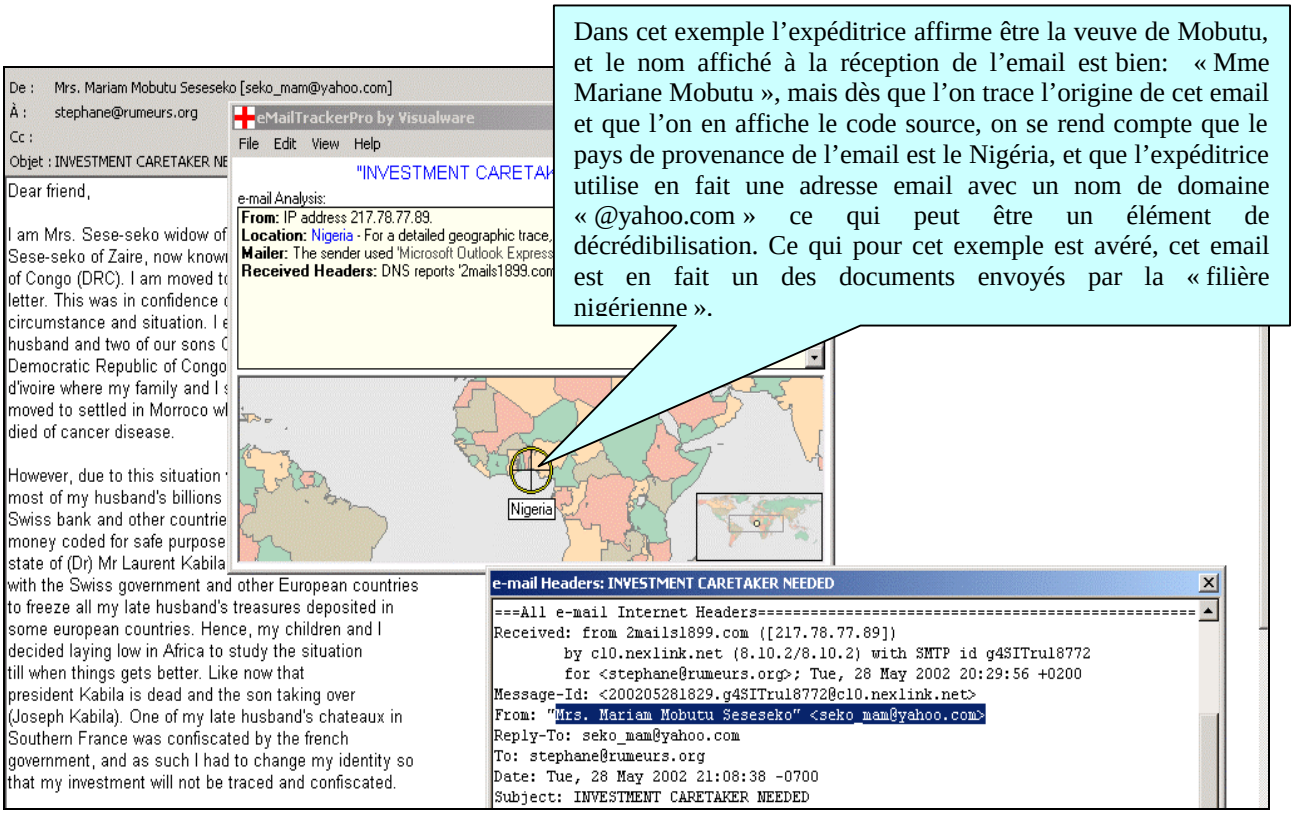
7.3) Crédibilité de l'information : les démarches de validation « techniques »

- **provenance de l'information** *: quel est le nom de domaine du site visé, est-ce un nom courant, provient-il d'un pays au bénéfice d'une législation permissive, quels sont les pré requis pour l'enregistrement d'un nom de domaine (dans le cas d'un nom par pays) ? On peut vérifier les différentes procédures à l'adresse suivante :* ***www.iana.org/cctld/cctld-whois.htm****. Pour les noms de domaines génériques se rendre sur :* ***www.iana.org/gtld/gtld.htm****)Pour vérifier l'existence d'un nom au niveau mondial se rendre sur :* ***www.uwhois.com/cgi/domains.cgi?User=NoAds***
- **source, hébergement et traçabilité d'un site Web** *: les autorités de régulation au niveau de la distribution des adressage régional des adresse IP sont listées à l'adresse suivante :* ***www.iana.org/ipaddress/ip-addresses.htm****, de même que l'on pourra s'intéresser à l'attribution des classes d'adresses IP sur ce lien :* ***www.iana.org/assignments/ipv4-address-space.*** *Des produits logiciels tels que « Visualroute » de la maison « Visulaware » peuvent apporter une aide précieuse au non-technicien pour la tracabilité d'un site et le l'identité de l'hébergeur :*



- **l'identification de l'ayant droit économique (détenteur d'un site Web)** *: pour identifier le détenteur d'un site web on va devoir dans un premier temps se rendre sur :* ***www.internic.org/whois.html*** *afin de savoir auprès de quelle société le nom de domaine a été loué. En effet depuis la libération à la concurrence de la location des noms de domaines un certain nombre de sociétés d'enregistrement (Registrars) ont vu le jours. L'url susmentionnée va nous permettre de trouver le Registrar concerné, ainsi on pourra se rendre sur le site de celui-ci et utilisé l'outil (Whois) prévu à cet effet pour identifier le détenteur de nom de domaine faisant l'objet de notre recherche. On peut donc constater qu'il n'y a pas de relation entre les différents « Whois » qui pourtant enregistrent le même type d'extensions (.biz, .com, .org, .info,etc : voir annexe pour accéder à toutes les extensions.). Le fait que l'on a identifié le détenteur du site, ne signifie pas que l'on a les informations sur l'hébergeur du contenu du site (celui-ce sera trouvé au moyen de Visualroute, tel que démontré dans l'exemple figurant au paragraphe précédent)*

- **source et traçabilité d'un email:** *« email traker pro » de la maison « Visulaware »*



code html d'une page Web : Dans l'exemple ci-contre, l'on peut remarquer que « *CONTENT= -Intrusion par Kain-* » figure dans les en-tête méta de la page web. Ce qui signifie qu'il ne sera pas visible lorsque l'on visitera la page en question avec son navigateur. Certains moteurs de recherches sont capables de faire des requêtes dans la zone des métas (par exemple www.voila.fr, avec les recherches approfondies)

7.4) Remarque concernant l'identification des détenteurs d'un site Internet

Quelles est la problématique au niveau légale ?

Le principal problème que l'on rencontre se situe au niveau de la nécessité de l'entraide judiciaire et du manque de contrôle de la crédibilité des informations fournie par les ADE (ayants droits économiques), de plus la possibilité de séparer les différentes informations en relation avec les détenteurs de sites (au niveau géographique humain et des prestataires de services), Par exemple récemment un cas d'escroquerie avait pour contexte un nom de domaine Thaïlandais, un contenu hébergé en Californie, et l'adresse du détenteur située en Espagne.

Exemple d'identification transnationale :

Propriétaire du nom de domaine *SXXXXXX-XXXXX.COM* Adresse postale : **1) Rxxxxx Mxxxxx & Co Limited** Rxxxxx Mxxxxxx & Co Limited Mr Rxxxxx Mxxxxxx Kxxxxx Gxxxx Road LE2 2LF Leicester GB	Location du nom de domaine *SXXXXXX-XXXXX.COM* **2) Schlund + Partner AG** C'est la société d'enregistrement [Registrar] par laquelle M. Matisson est passé pour enregistrer le nom de domaine *SXXXXXXX-XXXXX.COM.* Adresse postale : Schlund + Partner AG Erbprinzenstraße 4 – 12 76133 Karlsruhe Germany
Hébergement du contenu en relation avec *SXXXXXX-XXXXX.COM* **3) Cyberporte** Cyberporte héberge le site, ou contenu du nom de domaine *SXXXXXX-XXXXX.COM*, elle loue l'espace (hébergement des données) nécessaire à fournir ses prestations de service chez la société anglaise *WEBFUSION*, elle-même filiale de la société *HOSTEUROPE*, celle-ci a aussi la gestion des serveurs de nom qui hébergent le nom de domaine *SXXXXXX-XXXXX.COM* Adresse postale : Hxxxx, Mxxxx 2 mxxxxx des Cxxxxxx Lauris, 84360 FR	Hébergement du nom de domaine *SXXXXXX-XXXXX.COM* par l'intermédiaire de *CYBERPORTE* **4) hosteurope.com** C'est la société [Registration Service Provider] qui à mis à disposition les serveurs de noms [voir ci-dessous : nserver] nécessaires à la prise en charge du nom de domaine *SXXXXXX-XXXXX.COM* sur l'Internet. *HOSTEUROPE*, par l'intermédiaire de l'une de ses filiales anglaises : *WEBFUSION* [http://www.webfusion.co.uk/corpinfo.shtml], n'étant que le prestataire de service de *CYBERPORTE* et n'a pas pour client direct *MXXXXX & CO LIMITED* Adresse postale : Host Europe PLC Kendal Avenue London W3 0XA GB

8) Analyse de l'environnement et de la survenance de l'information par l'interprétation des signaux faibles

La notion de l'environnement de l'information peut se référer à la vérification du contexte dans lequel une information est diffusée (contexte alarmiste, tendu, favorable ou défavorable) est-ce que la teneur de l'information a une influence concrète sur une situation actuelle. La survenance quant à elle fait référence au moment ou l'information apparaît, et l'influence qu'elle a sur les éléments présents et à venir, ainsi que les facteurs de coïncidences informationnelles que l'on peut discerner.
Pour améliorer l'efficacité de cette technique on utilisera la méthode « PUZZLE » (développée par le professeur Humbert Lesca de l'université Grenoble, elle est basée sur l'analyse des signaux faibles). Il s'agit de puiser dans diverses sources des éléments d'information de type heuristiques (des brides d'information) pour ensuite les regrouper sur un même niveau d'analyse afin d'établir les « liens relationnels » qui seraient susceptibles d'exister entre les différents éléments pré-sélectionnés (de causalité ou de contradiction par exemple).

8.1) Modèle d'analyse de l'environnement de l'information.

a) Quel est le fait ?

b) **Identifier la source** (*notion* d'*environnement de l'information*)

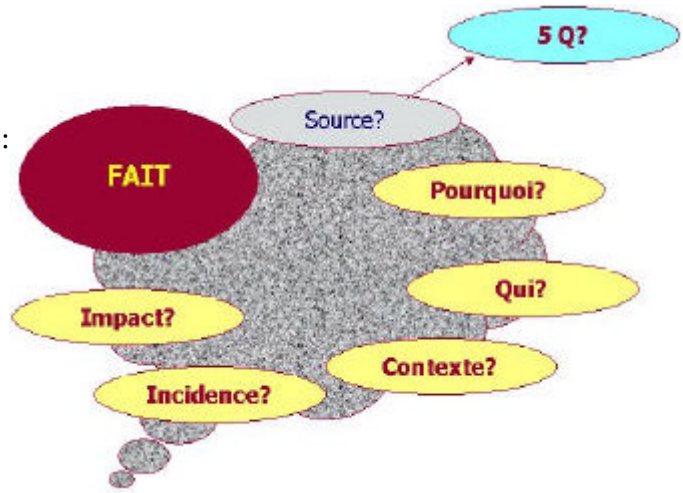
- *Crédibilité (média, auteur);*
- *Environnement (politique,sociale);*
- *Localisation (pays).*

c) **Les 5Questions en relation avec la source** :

- *qui,*
- *quand,*
- *quoi ,*
- *à quel moment,*
- *pourquoi ?*

d) **Les 5Questions en relation avec le fait** :

- *Qui ?*
- *Pourquoi ?*
- *Contexte ?*
- *Impact ?*
- *Incidence ?*



8.2) Deuxième axe d'analyse du fait:

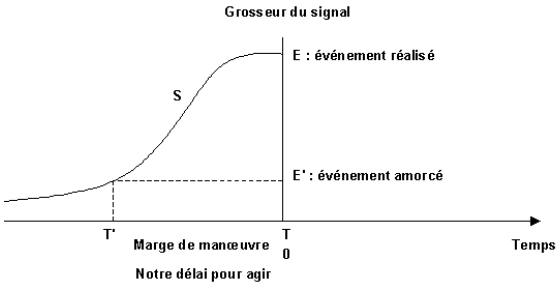
- **Les sources:** *indépendance vis à vis du journal ou d'autres acteurs impliqués. Couleur politique. Précision.*
- **Les faits:** *distinguer les faits, des opinions, des suppositions, des commentaires, des hypothèses, des supputations*
- **Les contradictions:** *recoupement, correspondance entre les infos. Détection des indices de contradiction entre les différents vecteurs de l'information en question*
- **Le débat:** *à qui donne-t'on la parole, qui est concerné, qui manque à l'appel…*
- **Les mots:** *quelle est la dialectique, comment sont employés les mots, charge émotionnelle, double sens.*
- **Le titre:** *reflète-t'il vraiment le contenu de l'article ou les faits annoncés, induit-il une interprétation erronée du contenu…*

- **Les chiffres:** *sont-ils précis, les sources sont-elles fiables, le contexte dans lequel ils sont utilisés correspond-t'il.*

8.3) Utilité et compréhension du modèle d'analyse des signaux faibles

Dans un contexte de surabondance de l'information, l'enjeu est de pouvoir distinguer parmi le "**bruit**" (masse d'information) l'information qui sera utile à l'entreprise. Il s'agit donc de détecter les faibles occurrences, c'est-à-dire les "**signaux faibles*** " L'idée de "signaux faibles" peut être définie à partir de la notion de "signaux d'alerte" (encore dénommée "signaux précoces") qui désigne le plus souvent des signaux de faible intensité. *I. Ansoff

Au niveau de l'analyse, on peut considérer les signaux faibles comme des brides d'informations, qui analysées séparément, ne signifient pas concrètement un événement à venir mais plutôt l'indice d'une situation potentiellement possible. C'est l'utilisation de faisceaux d'indices (hétérogènes) qui permettra de définir la probabilité du degré de réalisation de l'événement à venir. Cette approche « logique » d'anticipation peut-être formalisée et optimisée par le biais de l'intelligence collaborative.



source : Humbert Lesca

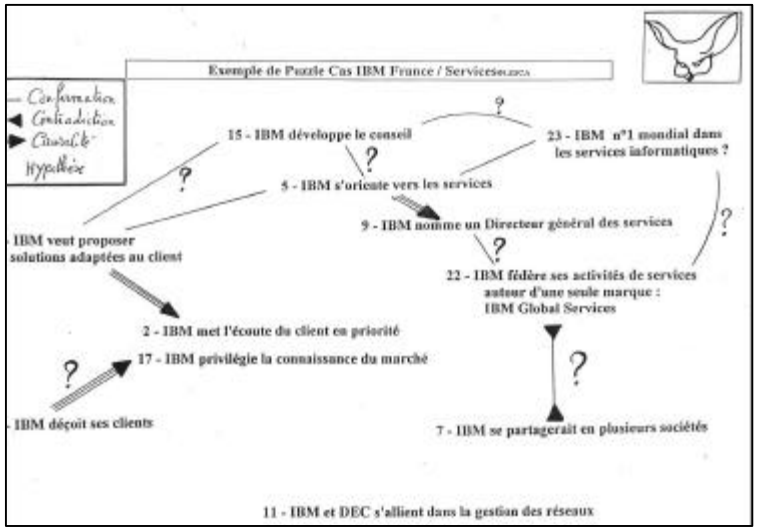
*A l'instant **T**, l'événement **E** est totalement réalisé.* ***Le signal S qui lui correspond est un "signal fort"*** *donc facilement perceptible. A ce moment là, le délai pour réagir à l'événement est nul,* ***il n'y a aucune marge de manœuvre possible****.*

*A l'instant **T'**, l'événement (à l'état **E'**) est juste amorcé. Le signal qui lui correspond est un signal faible, difficilement perceptible.* **L'effort à réaliser pour détecter ce signal est donc plus important**. En revanche, la **marge de manœuvre**, mesurée par **l'écart entre T' et T,** est alors suffisante pour faire face à la survenance de l'événement.

8.4) La méthode « PUZZLE » d'analyse des signaux faibles

Méthode Puzzle: le processus

1. Catégorisation des informations
2. Numérotation des informations
3. Définition des relations entre les informations, selon les liens de:
   - *Causalité*
   - *Confirmation*
   - *Contradiction*
   - *Hypothèse*
   - *Fréquence*



Les programmes tels que **I2**(analyser : www.i2.co.uk) **Watson** () ou **Mind Map** (www.mindmap.com), permettent aussi à différents niveaux, de travailler sur une approche qui regroupe : l'analyse de l'environnement de l'information, l'analyse des signaux faible et la méthode PUZZLE. Pour pouvoir « fusionner les méthodes susmentionnées on va partir du principe de fonctionnement du « data mining ». On va donc collecter un maximum d'informations d'environnement hétérogène tant au niveau de leurs sources, que de leurs catégories, que de leur crédibilité (à ce titre on annotera les informations selon un facteur de crédibilité, situé entre 1 et 10 par exemple). Au final on va essayer de procéder à l'agrégation et à la consolidation des éléments présents afin de procéder au traitement méthodologique des données

Rappel des démarches de bases

1) Collecte

- Identifier la source (notion environnement de l'information)
- Analyser la source:
- Crédibilité (média, auteur);
- Environnement (politique, sociale);
- Localisation (pays).
- Croisement des sources

2) Observation

- Observer le texte
- Sortir les éléments d'intérêt
- Chercher des informations complémentaires dans les sources blanches ou grises
- Numéroter les éléments définitifs selon deux méthodes: *chiffres* pour les éléments du texte, *alphabet* pour les éléments des sources.
- Croisement de l'information

3) Qualification et consolidation de l'information

- Principes de l'intelligence collaborative:
- Profils: cognitifs heuristiques et analytiques
- Groupe composé de savoirs hétérogènes
- Mise en circulation de l'information au sein du groupe – validation des informations par sa confrontation aux différentes connaissances présentes au sein du groupe
- Création de rapports d'étonnements

4) Schématisation :

- Développer les éléments selon la relation qu'il est possible de définir entre les différentes informations présentes;
- Analyse des signaux faibles
- Schématiser les différentes relations identifiées par un graphique;
- Développer des hypothèses selon des scénarios exploratoires (identiques à ceux que l'on utilise pour définir sa stratégie de recherche).
- Détection des indices d'étonnement
- Création de savoirs

8.5) Traitement de l'information : Principes de l'intelligence collaborative:

Profils: cognitifs heuristiques et analytiques : il est important de tenir compte des profils psychologiques présents dans la chaîne de traitement de l'information. Si l'on ne place pas les personnes au bon endroit (que cela soit au niveau de la mise en place d'une structure de veille ou lors d'un travail d'analyse de groupe). Un profil cognitif inductif sera plus à même de travailler avec des informations incomplètes, et donc aura plus de chance de trouver des indices lors d'une approche de détection des signaux faibles. Le profil cognitif analytique ne se contente pas d'information tronquée, il lui faut des données précises, on le placera donc en aval de chaîne de traitement de l'information (en partant de la collecte, par rapport au profil inductif).

Groupe composé de savoirs hétérogènes : dans le concept de l'intelligence collaborative il est important de bénéficier d'une diversité de savoirs. De même qu'il est positif d'avoir des savoirs atypiques au sein du groupe. Dans la phase de mise en circulation de l'information, la validation des informations par sa confrontation aux différentes connaissances présentes au sein du groupe. L'intelligence collaborative est d'autant plus efficace quand elle s'appuie sur des modèles tels que ceux abordés aux points 8.1 à 8.4.

8.6) Exemples d'utilisation des méthodes traités dans le chapitre 8

1er cas : analyse de l'actualité : Le « sniper » de Washington, on avait fortement suggéré à l'époque des faits (octobre 2002) que ce tueur pouvait avoir une relation avec le groupe terroriste d'Al Qaida. Dans un premier temps lorsque le tireur est apparu, on a traiter le premier meurtre comme un fait divers, par la suite avec la multiplication des homicides on a parlé d'un tueur en série. Le climat de psychose qui a commencer à s'instaurer et le fait que les autorités étaient impuissantes en face à ce criminel à poussé celles-ci à chercher un bouc émissaire. C'est à ce moment que l'analyse de l'environnement de l'information et de la survenance de l'information peut être exemplifiée.

Premier élément : Les autorités ont subitement demandé l'autorisation à Donald Rumsfeld d'employer des moyens militaires pour combattre « cet ennemi » à l'origine de la psychose de la population de tout un état (il s'agissait d'employer des drônes ainsi que des satellites).

Deuxième élément : les télévisions ont commencé à diffuser des reportages sur les effets psychologiques du terrorisme (documentaire sur les victimes d'attentats en Israël, etc…)

Troisième élément : La conseiller du président pour la politique de sécurité intérieure a affirmé qu'elle envisageait d'interroger les prisonniers de Guantanamo afin de vérifier s'il existait un lien entre le « sniper » et les terroristes

On peut tout à fait déterminer la volonté de création d'un lien implicite entre un événement d'actualité et l'utilisation de la psychose terroriste du moment par l'analyse de ces éléments de dissonances dans le contexte informationnel

2ème cas : les images montrant des palestiniens en liesses après les attentats du 11 septembre. Cet exemple peut permettre d'étudier, les aspects de coïncidence informationnelle ainsi que de survenance de l'événement. Si on analyse le contexte dans lequel ces images sont apparues sur les chaînes de télévisions, on constate que dans un premier temps elles avaient « leur place » dans la chaîne (contexte) d'information. Mais au niveau de la coïncidence et de la survenance on avait les premiers signaux faibles

Premier élément : alors que le monde est occupé, à New York, par la chute des tours, une équipe de cameramen a réussi à se trouver au bon endroit au « mauvais moment ».

Deuxième élément : le groupe de palestiniens filmé n'a jamais fait l'objet de plans reculés, les cameramen sont en fait restés concentrés sur un petit groupe de personnes – la situation n'était donc

pas représentative d'un mouvement général.

Troisième élément : les images sont « arrivées » très vite sur les chaînes de télévisions, et dans la majeur parti des diffusions, elles n'ont pas été accompagnées d'un commentaire explicatif.

Quatrième élément : il y a un plus de 18 mois, un journal israélien (Haaretz) proche de l'armée avait fait paraître un article sur la volonté de l'armée de se doter « d'une force d'intervention médiatique rapide » dont le but était de mettre en place un « outil » pour être plus à même de se battre sur le terrain de l'information.

Cinquième élément : les images ont été filmées par l'antenne israélienne d'un réseau de broadcast international

Tous ces points permettent de remettre en cause l'information reçue et de mieux détecter les messages induits (les Palestiniens se réjouissent de l'attentat… Les Palestiniens ne sont peut-être pas étrangers à cet attentat… Les Palestiniens sont des terroristes). Ce qui fait que, par exemple, quand le Premier ministre israélien a ordonné aux chars de pénétrer dans les territoires occupés, le taux de protestations est resté très faible.

Ces différentes méthodes d'analyse (chapitre 8) sont les seules parades à cet aspect moderne de gestion de l'information. Pour un cas tels que ceux du type CNN (chapitre 7.1), cela peut permettre de fournir les éléments nécessaires à une détection appropriée des informations erronées

9) Stratégies visant à la maîtrise des Flux informationnels

Le développement de structures ou cellules de veille stratégique est intégré dans la stratégie de management de l'information. Les divers éléments présentés au chapitre 3 dans le paragraphe «*Le risque informationnel »* , sont représentatifs de la nécessité de délimitation du périmètre informationnel de l'entreprise.

Le principal instrument de travail est l'information et la tâche est de gérer le flux de cette information selon des principes de : réception ; interprétation ; dissémination; action.

De plus en plus le management de l'information en entreprise devient une nécessité, car il faut gérer :

- *Le flux des informations produites par l'entreprise pour elle-même*
- *Le flux des informations prélevées sur l'extérieur et utilisée par l'entreprise*
- *Le flux des informations produites par l'entreprise à destination de l'extérieur*

Ces différents flux se départagent en deux catégories :

- ***Les informations d'activités****, utiles à l'entreprise pour assurer son bon fonctionnement,*
- ***Les informations de convivialités****, permettent de vivre ensemble et en relation avec les autres et d'influer sur leurs comportements.*

*« On peut considérer par exemple qu'une campagne de désinformation agit selon le même principe qu'un virus informatique: **une information erronée qui s'appuie sur les ressources d'un système organisé dans le but de le déstabiliser**. »*

9.1) La veille stratégique: principes de base et possibilités d'utilisation

On va inclure dans la gestion de son périmètre informationnel la palette des risques et vulnérabilités en relation avec l'activité professionnelle de l'entreprise. Pour une banque par exemple on va inclure le nom des personnages politiquement exposés dans le périmètre de surveillance, à ce titre il

faut prêter attention au fait qu'inclure un risque dans le périmètre informationnel de son entreprise ne signifie par forcément traiter ce risque depuis son entreprise :

Considérant que chaque recherche d'information est une information. Il faudrait, pour le secteur bancaire, passer par des sortes de « proxy humain » et donc ne pas traiter la recherche d'information directement au sein de l'entreprise, mais passer par une entreprise tierce, bénéficiant de toutes les garanties de confidentialité, afin de réduire le risque au minimum. On peut très bien imaginer créer une structure indépendante de sa propre entreprise, mais sous le contrôle de celle-ci, dédiée à la gestion du risque.

9.2) Philosophie pour la mise en place d'une structure de veille

Pour mettre en place cette structure de veille on va en claquer le principe de fonctionnement au niveau humain sur « l'intelligence collaborative (chapitre 8.5) ». Ensuite on va essayer de répartir le travail en rapport avec les compétences des collaborateurs de l'entreprise concernée, de manière générale, la veille n'est pas une occupation à temps plein. Une fois que l'on a pris en compte les aspects de personnalité, il est indispensable d'y marier les compétences nécessaires à la détection des signaux (domaine professionnel). On essayera autant que possible d'automatiser par secteur la collecte d'information, le premier tri devant être fait par les ressources du secteur concerné.

9.3) Etapes du cadre méthodologique d'une surveillance électronique:

Construire la liste des mots clés qui délimitent le périmètre de surveillance : En fonction des thèmes de surveillance, il est capital de construire une liste de mots clés en plusieurs langues qui seront la base des premières recherches manuelles.

Tester ses mots clés sur les moteurs de recherche : Evaluer le volume d'information existant sur le sujet et de définir plus précisément les expressions (bouts de phrase) qui donneront les résultats les plus précis.

Les recherches effectuées en texte intégral, si tous les mots clés ne sont pas utilisés dans les recherches, on risque de passer à coté de documents pertinents et qui contiennent d'autres mots clés que ceux de notre liste. Attention à la construction de la liste des mots clés et des expressions de recherche (" style Internet ").

Construire un carnet d'adresses des sites ou des pages à surveiller. En complétant les recherches sur les moteurs,

Hiérarchiser les sites à surveiller. Les sites ne sont pas à surveiller avec la même fréquence.

Sélectionner les outils nécessaires (logiciels, outils online, délégation de services) afin de pouvoir automatiser le maximum de points à surveiller pour être à même de consacrer toute l'attention nécessaire aux indices d'information.

Une petite parenthèse sur Factiva : ce type de base de données professionnelles comporte un ensemble de systèmes d'alertes et d'options de configuration très utiles. Mais il faut être conscient que cela ne suffit pas. Bien que l'information de Factiva soit qualifiée et provienne de plusieurs milliers de sources officielles, elle ne permet pas d'accéder à tout le périmètre des informations non officielles au sein desquelles on a un fort potentiel de détection de signaux faibles et autres indices d'information. De plus, les forums de discussions ne sont pas non plus pris en compte.

9.4) Actions défensives et préventives

La protection de l'image, exemple : On peut estimer le préjudice à l'image causé par le détournement de sa page d'accueil (appelé « defacement » ou « défacement ») entre le moment « T » ou l'infraction a été commise et le moment « T1 » qui correspond au retour à la normalité.

L'espace temps compris entre « T » et « T1 » multiplié par le nombre de visiteurs qui habituellement fréquente le site dans cette tranche horaire donne le potentiel de préjudice. En effet si l'on a mille visiteurs dans la tranche horaire en question cela signifie qu'un millier de personnes auraient été en mesure de faire une capture de la page Web détournée et de l'envoyer au travers de forum de discussions, au concurrents ou aux représentants de la presse. Il faut aussi imaginer que la « dimension » de l'entreprise sur le Web correspond à la taille de son écran, de ce fait si l'on « tag » la page en question, c'est comme si on arrivait à « taguer » tout le bâtiment dans la réalité. *Pour pouvoir répondre à ce risque on mettre en place une solution de veille automatique sur l'information de son propre site Web, de la sorte que si elle venait à changer le responsable en serait alerté dans les plus brefs délais (une solution produit performante est WebSite Watcher : www. aignes.com/)*

Les personnages politiquement exposés : en prenant en compte les nécessités de sécurités abordées au chapitre 9.1 ainsi que les aspects techniques du chapitre 9.4, on va utiliser le principe de veille sur un certain nombre de sites et de portail d'informations. On peut aussi faire une recherche des interlocuteurs potentiels dans les régions ou le Web ne recense pas encore les sources d'information (presse locale, rumeurs)

Les contres sites : par rapport au nom et au domaine d'activité de son entreprise il serait judicieux de faire une étude de risque afin de savoir quel est le danger d'une exploitation à mauvais escient d'un nom de domaine en relation avec son activité professionnelle. Il est en effet moins coûteux de louer une série de noms de domaine (en moyenne CHF 18.- / an) que de devoir intervenir par l'intermédiaire d'un avocat. Pour palier à ce risque, il faudra non seulement prêter attention aux différentes extension (pays ou autres domaines génériques tels que les .biz, ou .info) mais aussi aux contre sites potentiellement utilisables (www.jeboycottedanone.com, syz.com, www.bcgefraude.ch)

La marque : la marque peut-être susceptible d'être attaquée par le biais des méta-tags ou au niveau du « *positionsquatting* » (pour ce qui est du *cybersquatting* ou utilisation indue d'un nom de domaine dans un but spéculatif, l'OMPI a mis en place une procédure d'arbitrage)

Le positionsquatting est le fait de payer pour apparaître dans les premiers résultats lors d'une recherche sur une marque dont on ne détient pas les droits. Des recherches sur plus de 60% des entreprises du CAC 40 amènent vers des sites non officiels, qui ne détiennent aucun droit sur la marque. Dans ce cadre, on observe de nombreux cas de parasitisme car les entreprises les moins scrupuleuses ont acheté des positionnements sur les recherches sur des entreprises concurrentes pour détourner leur trafic. (Source : Raphaël Richard CVFM).

Les métas-tags : grâce aux fonctions avancées de certains MR tel que « www.voilà.fr » on peut faire de recherches afin de savoir si sa marque figure dans les métas-tag d'un concurrent ou d'un contre-site

Les en-têtes méta

```
<base href=" http://www.site.com" />
<meta name="author" content=" auteur" />
<meta name="description" content=« descriptif">
<meta name="description" content=« description de l'activité" />
<meta name="description" content=« nom de la société " />
<meta name="keywords" content=« mot clés">
<meta name="Copyright" content="société">
<meta name="robots" content="index,follow">
<link rev="made" href="mailto:webmaster@site.com" />
<link rel="top" href="http://www.site.com" />
```

exemple des possibilités de recherches approfondies de « www.voila.fr »

Les rumeurs : comme cela à souvent été souligné dans ce mémoire les fausses informations sont présentes en masse sur le Web. Un des vecteurs à la mode est l'email, il sert de support à tous types de fausses nouvelles avec plus ou moins de succès. Profitant çà et là de la crédulité des internaute pour une part et de l'utilisation de la connaissance du comportement humain pour l'autre. C'est ce dernier point qui va être aborder dans l'exemple ci-dessous :

BONJOUR A TOUS,

UN DE MES CORRESPONDANTS A ETE INFECTE PAR UN VIRUS QUI CIRCULE SUR LE MSN Messenger. LE NOM DU VIRUS EST jdbgmgr.exe L'ICONE EST UN PETIT OURSON. IL EST TRANSMIS AUTOMATIQUEMENT PAR MESSENGER AINSI QUE PAR LE CARNET D'ADRESSES. LE VIRUS N'EST PAS DETECTE PAR McAFEE OU NORTON ET RESTE EN SOMMEIL PENDANT 14 JOURS AVANT DE S'ATTAQUER AU DISQUE DUR. IL PEUT DETRUIRE TOUT LE SYSTEME.

JE VIENS DE LE TROUVER SUR MON DISQUE DUR! !! AGISSEZ DONC TRES VITE POUR l'ELIMINER COMME SUIT:
1; Aller à DEMARRER, faire "RECHERCHER"
2. dans la fênetre FICHIERS-DOSSIERS taper le nom du virus:jdbgmgr.exe
3. Assurez vous de faire la recherche sur votre disque dur "C"
4. Appuyer sur "RECHERCHER MAINTENANT"
5. Si vous trouvez le virus L'ICONE EST UN PETIT OURSON son nom jdbgmgr.exe
" NE L'OUVREZ SURTOUT PAS!!!!!
6.Appuyer sur le bouton droi de la souris pour l'eliminer (aller à la CORBEILLE) vous pouvez aussi l'effacer en appuyant sur SHIFT DELETE afin qu'il ne reste pas dans la corbeille.
7. aller à la CORBEILLE et l'effacer definitivement ou bien vider la corbeille.

SI VOUS TROUVEZ LE VIRUS SUR VOTRE DISQUE DUR ENVOYEZ CE MESSAGE A TOUS VOS CORRESPONDANTS FIGURANT SUR VOTRE CARNET D'ADRESSE CAR JE NE SAIS PAS DEPUIS QUAND IL EST PASSE.DESOLEE POUR CET INCIDENT! ET MERCI D'AGIR VITE

Ce message est très bien construit, car si l'on suit la démarche décrite ci-dessus à la lettre on obtiendra en toute logique le résultat annoncé. Le seul problème est que le fichier qui est mis en cause et accusé d'être un virus destructeur, n'est que l'un des nombreux fichiers utiles au fonctionnement de l'ordinateur. Il était donc normale que les programmes anti-virus n'y prêtent aucune attention…

10) Conclusions

L'information, en tant qu'entité à proprement parlé, à pris dans un espace temps restreint une importance croissante au niveau de sa capacité de nuisance. Les nouvelle contraintes engendrées par cet état de fait ne peuvent plus êtres négligées. On a pu constater que la survie d'une entreprise peut être menacée par un nombre important des cas présentés dans ce mémoire, et que leurs mises en œuvre ne nécessitent pas l'engagement de beaucoup de moyens. La délimitation « des frontière du risque » s'est aussi étendue de manière considérable alors que les protections légales ont en été rendues plus difficiles d'accès et plus coûteuses. Les entreprises en phases avec des réalités à court terme ont de la peine à prendre en compte ce type de risque, la notion d'incertitude les poussent à faire le minimum pour assurer leur pérennité de leur activité professionnelle. Les données actuelles démontrent que malgré un fragilisation de fonctionnement due à l'utilisation des SI et à un « turn over » important elles préfèrent continuer à naviguer à vue. Il est vraisemblable que l'arrivée d'une responsabilisation pénale des dirigeants insouciants remettra au goût du jour la culture de la prévention du risque.

ANNEXE I : Glossaire des Termes Internet

Adresse IP(Adresse Internet Protocol)
Adresse unique permettant d'identifier un ordinateur sur l'Internet.

Applet
Une "applet" est un petit programme écrit en Java qui s'insère dans les pages HTML. Ce programme est ensuite interprété et exécuté par le navigateur.

Archie
Système qui permet de localiser un nom de fichier dans les sites FTP anonymes. Les serveurs Archie répertorient des millions de fichiers qui se trouvent dans quelques milliers de sites FTP anonymes dans le monde.

ARP (Address Resolution Protocol)
Protocole de Résolution d'Adresse. Messages et procédures utilisé par tout protocole de communication pour retrouver à partir des adresses locales, les adresses réseau. Dans TCP/IP, le protocole pour convertir les adresses IP réseau et les adresses physiques.

ARPANET (Advanced Research Project Agency Network)
Réseau à Communication par paquets qui constitua la base du réseau Internet. Ce réseau a vu le jour en 1969.

Authentification
Vérification de l'identité prétendue d'un ordinateur ou de l'utilisateur d'un réseau informatique.

ASCII (American Standard Code for Information Interchange)
Code binaire permettant de représenter les différents caractères. C'est le mode utilisé par presque tous les micro-ordinateurs pour coder les caractères saisis.

Autoroutes de l'information
Réseaux de télécommunications permettant la diffusion d'informations en mode numérique de façon aussi accessible que le téléphone ou la télévision, c'est-à-dire autant à partir des habitations que des lieux de travail. L 'autoroute de l'information (ou les autoroutes...) implique la convergence de l'informatique et des télécommunications sur les plans techniques et économiques et de grandes possibilités de diffusion d'informations variées, en mode texte, graphique, vidéo et son..

Bande passante
Gamme de fréquences qu'un instrument peut produire ou qu'un canal peut transmettre sans affaiblissement du signal. La largeur de bande s'exprime en Hertz. Plus la Bande Passante d'un réseau est élevée, plus grande est son aptitude à transmettre un flot important d'informations.

Backbone
Epine dorsale d'un réseau et point de concentration de celui ci. Ce terme peut être employé pour l'infrastructure ou pour les services (comme la diffusion de News) Il désigne une ligne haute vitesse ou un ensemble de lignes haute vitesse qui constitue un point de passage important dans un réseau. Ce peut être également une machine servant de point de concentration. C'est en fait un nœud de communication..

BBS (Bulletin Board System)
Messagerie entre micro-ordinateurs abonnés, permettant la mise en place de forums et de dialogues en direct. A une échelle plus grande les BBS sont appelés "services en ligne" ("On Line Services'? Parmi les exemples de "services en ligne" citons Compuserve, America On Line(AOL).

CACHE
Dispositif matériel éventuellement associé à un composant logiciel dont l'objectif est de stocker localement des ressources afin de diminuer le délai de mise à disposition de ces ressources. Il peut s'agir de mémoire dite "mémoire cache" associé à un microprocesseur et dont l'objectif est de stocker dans le microprocesseur des données afin d'éviter une perte de temps à aller chercher ses données sur un support extérieur (mémoire ou disque). On parle alors d'antémémoire.
Le mécanisme de cache peut-être aussi mis en oeuvre sur un serveur pour stocker temporairement des données fréquemment utilisées et se trouvant sur un site distant (Proxy Server). C'est aussi un mécanisme utilisé par certains protocoles comme le DNS ou ARP.

Chiffrement

Méthode de protection des données. Lorsqu'on accède à ces données, elle empêche de les comprendre sans le recours d'une clé de chiffrement permettant d'afficher "en clair".

Clé publique
Clé utilisée dans un système de chiffrement dans lequel la clé de chiffrement est différente de la clé de déchiffrement. Ce système repose sur le caractère secret d'une clé dite privée même en connaissant la clé publique (c'est-à-dire diffusée publiquement). Un modèle de chiffrement à double clé est celui de POP ("Pretty Good Privacy").

Client-serveur
Mode de fonctionnement d'un programme informatique qui répartit la charge de travail d'une application entre deux logiciels: le client et le serveur. Le client assume les échanges avec l'utilisateur, la préparation des requêtes, l'affichage des résultats, etc. Le logiciel serveur assume la gestion des bases de données et effectue les traitements, les recherches, et traite tout type de requêtes que lui adresse le logiciel client.

Commutation par paquet
Mode de transmission des informations sur l'Internet: les données à transmettre sont découpées en plusieurs paquets et chaque paquet est envoyé de manière indépendante. Ce mode est à distinguer de la commutation de circuit qui est utilisée pour le téléphone.

Compression
Traitement des données numériques qui réduit leur volume. Dès lors que des informations sont numérisées (converties en séries de bits), elles peuvent être compressées afin d'occuper moins de place. Les techniques de compression impliquent un codage par algorithmes mathématiques. La décompression est ensuite effectuée grâce aux formules inverses. Ces techniques ont tant progressé que l'on peut à présent compresser des données, et les restituer sans perte de qualité, dans un rapport de 1 à 20.

Cookies
Fichier de type texte (.txt), qui s'inscrit sur le disque dur à l'aide de certaines pages Web, dans le but d'être réutilisé plus tard par d'autres pages Web.

Cryptage
Méthode qui assure la confidentialité et la sécurité de l'information véhiculée sur l'Internet. Les données sont brouillées, et donc illisibles, puis débrouillées à l'aide de la même méthode. Certains navigateurs Web, comme Netscape. Possèdent de telles fonctions.

Cybercafé
Café dans lequel on peut se connecter à l'Internet . Les premiers cafés de ce type sont apparus en 1992 à San Francisco.

Cyberspace
Terme de William Gibson, romancier, désignant les mondes virtuels constitués par les réseaux informatiques

Débit
Mesure la quantité d'informations que peut transmettre un canal de transmission pendant un temps donné Généralement en bits par seconde pour les transmissions numériques.

DNS (Domain Name System)
Base de données distribuée permettant de faire la correspondance entre nom de machine logique et l'adresse IP
E-mail (Electronic Mail)
Application qui permet à un utilisateur d'échanger des messages avec d'autres utilisateurs dotés d'une adresse Internet, communément appelée adresse de courrier électronique. L'Office de la langue française du Québec recommande l'usage du terme "messagerie". En France on préconise aujourd'hui l'usage de "Mél". Certains utilisent le terme plus heureux de "courriel ". On peut aussi entendre le terme BAL (boite aux lettres). La messagerie électronique fut l'un des tous premiers services du réseau Internet mis en place à partir de 1971. Comme la plupart des services développés sur Internet la messagerie électronique utilise un mécanisme client-serveur. Pour pouvoir envoyer un message, l'utilisateur doit disposer d'une boite aux lettres électronique (souvent liée à un compte sur une machine), comprenant généralement son nom et celui de la machine sur laquelle il est enregistré. L'adresse a la forme nom@machine, cette boite aux lettres lui est réservée. Le protocole utilisé sur Internet est SMTP.

Ethernet
Norme d'équipement utilisée dans les réseaux locaux. Ce type de réseau peut supporter le protocole TCP/IP utilisé par l'Internet à un débit nominal de 10 Mbps, très répandu dans le monde de la micro-informatique.

Forums de discussion ("news group")
Espaces de rencontre et de dialogue sur le Web. Ils sont classés par thèmes et par pays (donc par langue). Les sujets de conversation sont très variés, de l'aquariophilie au cinéma d'art et d'essai, en passant par la bande dessinée et la psychanalyse. Le principe de ces dialogues est simple et se rapproche de celui des Listes de Diffusion. Ils peuvent être modérés ou non. Lorsqu'ils le sont, les interventions des abonnés sont évaluées et déposées sur le forum, uniquement si elles sont effectivement en rapport avec le thème de celui-ci.

Fournisseur d'accès Internet
Société qui loue des connexions à l'Internet et fournit les services associés nécessaires. On trouve également souvent l'appellation ISP (Internet Service Provider).

FTP (File Transfer Protocol)
Ce service de !'Internet permet de télécharger des fichiers. Il permet aussi de déposer (télé verser) des fichiers dans un site donné.

Gopher
Système distribué d'accès à l'information conçu à l'Université du Minnesota en 1991. Très simple, on l'utilise surtout pour diffuser et consulter des documents. La présentation et la navigation s'effectuent à l'aide de menus. Il tend à disparaître au profit du Web. .

GUI
Graphical User Interface - Interface utilisateur graphique.

Home Page ou Page d'accueil
Nom donné au document principal d'un site Web. Ce document constitue le document racine de l'arborescence de la base de donnée du site. C'est aussi le nom donné au document de présentation d'un utilisateur. En général ce document présente le site, la société ou la personne. C'est le point de départ de la navigation dans la base de données du site.

Hôte (Host, Host computer, Host system)
Ordinateur hébergeant un service Internet, dont un ordinateur client peut importer les données et les informations en s'y connectant. Plusieurs services peuvent résider sur la même machine hôte (un serveur de courrier électronique et un serveur Web par exemple). Inversement, plusieurs hôtes peuvent se partager l'hébergement d'un même service, nécessitant d'importantes ressources machines.
Tout ordinateur d'un réseau mettant des services à la disposition des autres systèmes du réseau. Il propose notamment les services de messagerie et un serveur Web. Dans !' Internet, il supporte les protocoles TCP/IP et possède une adresse Internet.

HTML (HyperText Markup Language)
Langage de marquage de documents. Ce langage offre une présentation de l'information qui permet une lecture non linéaire grâce à la présence de liens sémantiques activables dans les documents. C'est un sous-ensemble de SGML.

HTTP (HyperText Transport Protocol)
Protocole de transmission de documents hypermédias. Il est utilisé pour transférer des documents hypertextes ou des documents hypermédias entre un serveur et un client W3.

Hypermédia
Ce mot est formé à partir de "hypertexte" et de " multimédia". Il caractérise l'ensemble des techniques offrant la possibilité de lire ou produire des documents numériques contenant du texte, de l'image ou du son en passant de l'un à l'autre par des liens hypertextes. Dans le W3, il s'agit de plus en plus d'hypermédia, bien que l'on parle souvent d'hypertexte.

Hypertexte
Présentation de l'information permettant une lecture non linéaire grâce à des liens sémantiques activables dans les documents.

Hytelnet
Base de données mise à jour régulièrement et constituée principalement de références à des sites Telnet et à d'autres sites Internet.

IMAP (Internet Message Access Protocol)
Protocole d'accès aux messages Internet permettant l'accès aux messages E-mail et BBS se trouvant sur un Serveur de messagerie. Le protocole IMAP est un protocole d'accès concurrent au protocole POP. Ces deux protocoles sont particulièrement utiles pour la lecture de son courrier à partir d'un poste nomade connecté au serveur par le biais d'un

réseau public de transmission. Contrairement à POP, IMAP permet de ne transférer que les entêtes des messages lors de la lecture de la boite aux lettres, tandis que POP transfère la totalité du contenu sur le poste local.

Interface
Désigne la frontière à travers laquelle deux systèmes communiquent. Une interface peut-être un connecteur matériel reliant deux équipements ou un ensemble de conventions utilisées par deux systèmes logiciels pour communiquer entre eux.

ISP (Internet Service Provider.)
Voir Fournisseur d'accès Internet

INTERNET
Réseau constitué par un ensemble de réseaux télématiques qui interconnectent la plupart des pays du monde. L'apport d'Internet par rapport à d'autres réseaux est d'être basé sur un protocole de communication TCP/IP indépendant du type de machine (Mac, PC, Unix,...), du système d'exploitation et du support de transport physique utilisé. De plus, Internet fonctionne de manière décentralisée: Son fonctionnement ne dépend ni d'administration ni d'ordinateur central. Un paquet d'informations peut aller d'un point à un autre en empruntant potentiellement plusieurs chemins.

Intranet
Ce terme représente l'utilisation des protocoles et services Internet dans les réseaux internes des entreprises.

IP (Internet Protocol, protocole Internet).
Le protocole réseau que toute machine sur l'Internet utilise pour communiquer avec une autre.

IRC (Internet Relay Chat)
Service disponible sur l'Internet permettant la discussion à plusieurs en temps réel avec d'autres personnes par échange de messages textuels (de l'anglais "to chat", bavarder).

LAN (Local Area Network)
Voir RESEAU LOCAL

Lien hypertexte
Zone activable d'un document hyper textuel permettant d'établir une connexion entre des données ayant une relation de complémentarité entre elles, et ce, où qu'elles se trouvent dans l'Internet. Les termes " pointeur" et "marqueur" indiquent respectivement le lien hypertexte (pointeur) et la zone activable (marqueur).

Listes de diffusion
Appelées aussi listes de distribution : Ce sont des listes auxquelles on peut s'abonner gratuitement et qui véhiculent des informations par thème.

Listserv
Application qui supporte les échanges entre un groupe de personnes qui possèdent une adresse électronique. Une liste peut être privée ou publique, modérée (i.e. être "gérée" par un modérateur) ou non modérée. Elle est utilisée pour les discussions, mais aussi pour la livraison de journaux électroniques. Majordomo est une application équivalente.

Login
Nom de connexion. Code d'accès unique qui identifie un utilisateur lorsqu'il accède à un ordinateur. C'est aussi l'opération qui permet cet accès à un ordinateur.

MAN (Metropolitan Area Network)
Réseau métropolitain dont la taille est située entre celle d'un réseau local LAN et celle d'un réseau longue distance WAN. Ce peut être un réseau à l'échelle d'un campus ou d'une ville.

Mime (Multipurpose Internet Mail Extension)
Standard utilisé par la messagerie pour coder des fichiers binaires (son, images, programmes). L 'apport principal de MIME est le support du format 8 bits permettant l'envoi direct de tout type de document. Plus précisément. MIME est défini dans le RFC 1341. Les extensions MIME ont été faites pour corriger les limitations initiales de la messagerie Internet telles que définies dans le RFC 822 datant de 1982. En particulier ces extensions ont été faites pour être indépendantes de la machine émettant, transmettant ou recevant le message. Elles permettent de préciser les attributs du message ou de certaines de ses parties comme le format et le type de contenu, le codage (7 bits, 8 bits, base 64...), mais aussi l'alphabet, la langue, la description..

Modérateur

Personne qui se charge de filtrer les articles diffusés sur les News Groups ou Groupes de nouvelles dites "modérées ".

Mosaic
C'est la première interface graphique ou Navigateur qui a permis l'accès à la plupart des applications qu'on retrouve dans le réseau Internet (www, Gopher, Telnet, FTP, News.) Des versions de Mosaic existent notamment en environnements X Window, Macintosh et Windows. Son auteur est un des fondateurs de Netscape.

Moteur de recherche
Outil de recherche d'information sur l'Internet. Ce terme est surtout utilisé pour la recherche dans le Web. Exemple Voilà, AltaVista, Yahoo,

Multimédia
Ensemble de techniques permettant d'utiliser des informations de type texte, image fixe, image animée et son sur un même support numérique et interactif.

Navigateur (browser)
Programme qui sert d'interface entre l'utilisateur et le réseau. Ex: "Netscape Navigator" ou "Internet Explorer".

News
Nouvelles Usenet : ce sont des Forums de Discussion où chacun dépose des courriers (articles) par thème. Ces courriers sont conservés quelques jours et donnent lieu à des discussions. Une hiérarchie dans l'organisation des groupes permet d'identifier ceux qui existent sur les différentes thématiques.

Newsgroup
A traduire par "Groupe de nouvelles" ou "Forum Usenet" utilisant généralement le réseau l'Internet, désigne un groupe de discussion sur un sujet particulier.

Niveau application
Niveau où une application, comme le courrier électronique, Web ou Gopher, se réalise. Ce protocole applicatif se situe au-dessus de la couche de transport de l'information.

NNTP (Network News Transfer Protocol)
Protocole utilisé par Usenet pour transférer des fichiers de News d'un serveur à l'autre.

Nom de domaine
Element d'une adresse électronique qui permet de la classer en fonction de la localisation, de l'activité ou du nom du propriétaire du domaine. Le top-level domain indique la localisation géographique du serveur lorsqu'il se compose de deux lettres (.fr pour la France,. uk pour l'Angleterre,. de pour l'allemagne, etc.).
Les entreprises ou les particuliers peuvent déposer un sous-domaine à leur nom propre. Dans l'acception courante, on considère que dans les adresses "http ://www.uunet.fr" et "info@uunet.fr", le nom de domaine est "uunet.fr".

On-line ou **Off-Iine**
Se dit d'un ordinateur ou d'un service lorsqu'il est connecté où déconnecté du réseau.

Opérateur Télécom
Désigne une société ou un organisme exploitant un grand réseau de télécommunications. Exemple: AT&T, France Télécom, Mercury...

Paquet
Petit ensemble de données faisant partie du transit d'une information à travers un protocole de commutation par paquet, comme TCP/IP.

Pare-feu (Firewall)
Dispositif matériel et/ou logiciel qui contrôle l'accès à l'ensemble des ordinateurs d'un réseau à partir d'un seul point d'entrée. Le firewall est en général situé entre le réseau interne et le monde extérieur, dans une zone appelée "zone démilitarisée".
La première fonctionnalité d'un garde barrière, est de filtrer les paquets qui transitent entre le réseau que l'on veut protéger et les réseaux extérieurs. Ainsi certains paquets peuvent être interdits de passage en fonction :
- de l'adresse de la source ou de la destination du paquet,
- du type de protocole (http, ftp, mail),
- du type d'applicatif,
- de l'heure et de la destination du paquet (accès interdit en dehors des heures ouvrables par exemple),

- A cette fonction basique de filtrage peuvent être associées des fonctions de sécurité avancées ; Telle la détection de virus, le masquage des adresses IP du réseau protégé ou encore l'établissement de tunnels cryptés associé à un procédé d'authentification.

Passerelle
Configuration matérielle ou logicielle assurant la communication entre deux protocoles distincts. Par exemple : dispositif assurant la communication entre un système de courrier électronique interne et le courrier électronique Internet. La passerelle effectue les traductions nécessaires pour que les données soient reconnues par les différents systèmes. Le terme passerelle (qui normalement ne désigne que les équipements effectuant une traduction des protocoles au niveau 7 et au-dessous) est souvent utilisé pour désigner des équipements spécifiques d'interconnexion comme les Routeurs.

Pointeur
Chaîne de caractères qui permet d'indiquer de manière unique la localisation d'une ressource. Un URL est un pointeur permettant d'accéder à une ressource du Web.

POP (Post Office Protocol)
Protocole d'accès au Bureau de Poste. Protocole permettant l'accès aux messages E-mail et BBS se trouvant sur un serveur de messagerie. Le protocole POP est un protocole d'accès concurrent au protocole IMAP. Ces deux protocoles sont particulièrement utiles pour la lecture de son courrier à partir d'un poste nomade connecté au serveur par le biais d'un réseau public de transmission. POP est plus ancien que IMAP et possède de moins riches fonctionnalités. Contrairement à POP, IMAP permet de ne transférer que les entêtes des messages lors de la lecture de la boite aux lettres, tandis que POP transfère la totalité du contenu sur le poste local.

Postmaster
C'est la personne qui, sur un serveur de messagerie, est responsable du bon fonctionnement du service. Il est le destinataire de tous les messages d'information de ce serveur. Il existe en général un compte "Postmaster" sur tous les bons serveurs de messagerie.

Proxy
Nom donné à un programme, une fonctionnalité ou à un serveur qui agit en tant qu'intermédiaire dans un échange d'information en effectuant un contrôle le plus souvent lié à la sécurité.
Voir les deux types de Proxy :
Proxy Gateway Type de dispositif pare-feu (Firewall) installé entre deux réseaux et qui protège les ordinateurs d'un réseau interne contre les accès des utilisateurs extérieurs. C'est en général un programme installé sur une Passerelle et qui bloque le passage direct des Paquets entre le client et le serveur et n'autorise le passage que de certains paquets. On parle aussi de relais applicatif, de machine bastion dans un sous-réseau démilitarisé. La plupart des Navigateurs peuvent être configurés pour utiliser les services d'une passerelle Proxy, c'est d'ailleurs dans certains réseaux sécurisés la seule façon pour accéder à des documents se trouvant à l'extérieur du réseau local (à condition encore qu'il existe aussi une passerelle avec l'Internet). Les navigateurs peuvent être configurés en fonction de la méthode d'accès (protocole) FTP, Gopher, Wais, News et HTTP.
Proxy Server Programme qui fournit un Cache pour des éléments présents sur d'autres serveurs qui sont soit présumés trop lents, soit éloignés ou coûteux d'accès. Ce terme est utilisé tout particulièrement dans le cadre du www. Un serveur qui reçoit une requête demandant un URL à l'extérieur : vérifie s'il n'a pas très récemment répondu à une requête identique.
- dans le cas où la page a été stockée dans son cache, il lui suffit alors d'extraire la page correspondante du cache pour la transmettre au client qui lui en a fait la demande. (Cela se traduit par un gain en temps de réponse, et éventuellement en coût si le transfert en provenance du serveur original se traduit par des dépenses réseaux particulières).
- dans le cas où la page n'est pas dans le cache, le serveur Proxy transmet la requête vers le serveur hébergeant l'URL demandé puis transmet le résultat de la requête de l'URL au demandeur.

Bien évidemment, le cache ne garde les documents qu'un temps déterminé, contrôlé par un algorithme en fonction de leur date d'entrée, taille et historique d'accès. La notion de serveur Proxy est à comparer à la notion de passerelle Proxy.

ppp
Point to Point Protocol. Protocole qui permet d'avoir accès aux fonctions du protocole IP à partir d'un modem et d'une ligne téléphonique conventionnelle. Le protocole Slip offre un service équivalent.

Protocole
Ensemble de règles qui définissent les modalités de fonctionnement d'une communication entre deux ordinateurs. Ou encore, méthode formelle de disposition des messages et des règles que doivent respecter obligatoirement deux ordinateurs ou plus pour échanger de tels messages.

Protocole ISO
Protocole dont les normes sont reconnues par l'ISO : International Standard Organisation (organisation qui s'occupe des standards au niveau international).

Real Audio
C'est une technique qui permet la transmission et le rendu de plages sonores sur Internet en temps réel.

RESEAU LOCAL (LAN -Local Area Network)
Système de communication mettant en relation permanente par des câbles plusieurs équipements informatiques (micro-ordinateurs, stations de travail, imprimantes et autres périphériques) à grande vitesse sur une courte distance (souvent un étage ou un immeuble au plus un ensemble de bâtiments situés sur un domaine privé). Il se définit par son système de câblage, sa vitesse, sa méthode d'accès et son logiciel de gestion. Les deux principales familles de réseaux locaux sont Ethernet et l'anneau à jeton (Token Ring).

RFC (Request for Comments)
Les RFC sont les documents servant à la définition de standards dans !'Internet. Il en existe aujourd'hui plus de 2000.

RNIS
Réseau Numérique à Intégration de Services. Réseau informatique et téléphonique Numéris qui offre des débits par canal de 64 Kbps. Il nécessite un abonnement particulier.

Routeur
Dispositif qui dirige vers un chemin ou un autre les paquets d'informations qui voyagent entre les réseaux. Il reçoit et retransmet des paquets de données entre différents segments d'un même réseau ou de réseaux différents.

Service en ligne
Service permettant d'accéder, par abonnement et à partir d'un ordinateur, à une information ou de réaliser une transaction à distance. Un service en ligne peut être accessible pour le particulier par l'intermédiaire de sa ligne téléphonique. Ces services offrent le plus souvent une passerelle vers l'Internet.

Serveur
Ordinateur relié au réseau et apparaissant comme fournisseur d'informations. Combinaison matérielle et logicielle assurant la prestation de services spécifiques à d'autres ordinateurs. Un seul serveur peut exploiter différents logiciels, offrant ainsi autant de services différents aux clients du réseau. Le client consommateur peut être un usager, un ordinateur ou un autre logiciel.

Serveur Web
Système informatique exécutant le logiciel qui permet d'accepter des requêtes utilisant le protocole d'application HTTP et servant à créer des sites Web ou à héberger des pages d'accueil personnalisées.

Shareware/Partagiciel
Logiciel utilisable à volonté selon les conditions énoncées en échange d'une somme d'argent versée à l'auteur. Il peut souvent être utilisé gratuitement pendant une période d'évaluation.

SGML (Standard Generalized Markup Language)
Norme la plus répandue de marquage de documents. HTML en est un sous-ensemble spécifique pour le marquage de documents hypertextes.

SLIP (Serial Line Internet Protocol)
Protocole qui permet d'avoir accès aux fonctions du protocole IP à partir d'un modem et d'une ligne téléphonique conventionnelle. Le protocole PPP offre un équivalant plus complet.

SMTP (Simple Mail Transfer Protocol)
Protocole utilisé pour échanger les messages entre les différents systèmes de messagerie qu'on retrouve sur les ordinateurs dans l'Internet.

TCP/IP (Transmission Control Protocol over Internet Protocol)
Né dans le milieu de la recherche militaire aux Etats-Unis, ensemble de protocoles qui rendent possible l'échange d'information entre une grande variété d'ordinateurs. Il repose sur la transmission par Paquet.

Télécharger
Effectuer le téléchargement d'une copie des données choisies par l'internaute d'un ordinateur à un autre en utilisant généralement le protocole FTP.

Telnet
Application qui supporte les sessions de connexion à distance en mode terminal à travers un réseau TCP/IP.

URL (Uniform Resource Locator)
Syntaxe utilisée par www pour spécifier la localisation physique d'un fichier ou d'une ressource sur l'Internet. C'est en quelque sorte le descripteur du chemin d'accès à une ressource du Web.

Usenet (Unix User Network)
Réseau des ordinateurs, transférant entre eux les fichiers de News. Usenet n'est pas l'Internet, même si aujourd'hui les deux réseaux sont fortement imbriqués.

VERONICA
Application qui permet de procéder à des recherches par mot-clé dans les menus des serveurs de type Gopher

Vérification
Saisie des données relatives à la sécurité d'un réseau. Les programmes de vérification servent à enregistrer les événements, à identifier les attaques du réseau et à s'assurer que le dispositif de sécurité du réseau fonctionne efficacement.

Visioconférence
La visioconférence est une technologie qui permet, depuis un micro-ordinateur, d'échanger avec un interlocuteur distant et de le voir en temps réel dans une fenêtre virtuelle à l'écran. Une application de cette technologie est le travail en commun sur des documents. Tout dispositif de visioconférence se compose d'une caméra vidéo, d'un microphone/écouteur et de cartes d'extension, pour la vidéo et la communication. Les échanges peuvent se dérouler point à point ou en mode multipoints.

Virus
Programme informatique parasite capable d'altérer parfois de façon irréversible le fonctionnement d'autres programmes. Les virus sont transmissibles par lecture de disquettes ou CD contaminés, et par communication en ligne.

WAIS (Wide Area Information Server)
Ensemble de logiciels qui permet de créer et d'interroger des bases de données indexées appelées bases Wais, et de rendre ces bases accessibles via l'Internet. Ce système supporte la recherche d'information en mode plein texte dans des banques de documents.

WAN (Wide Area Network)
En français : Réseau Longue Distance, c'est à dire qui va au-delà d'un site industriel ou commercial (dans ce cas on parle de LAN, au-delà d'un campus ou d'une ville (dans ce cas on parle le plus souvent de MAN. Les WAN font appel à l'infrastructure et aux services d'un ou plusieurs Opérateur Télécom et peuvent s'étendre sur plusieurs pays.

Web
En français, toile d'araignée : symbolise le réseau maillé de serveurs d'informations formant une toile d'araignée. Ces serveurs vont des pages personnelles aux interface s vers des bases de données. Par extension on parle de Web pour un serveur de documents HTML.

Webmaster
Nom attribué à une personne qui s'occupe d'un Serveur Web.

www (World Wide Web)
Système distribué d'accès à l'information qui s'appuie sur les principes de l'hypertexte et qui supporte les documents multimédias. Pour en savoir plus vous pouvez consulter le site du Consortium W3 : http://www.W3C.org Etant donné que le sigle WWW (en anglais) perd ses connotations culturelles en français, l'Office de la langue française recommande l'utilisation de "W3" pour des raisons phonétiques. Certains auteurs préfèrent l'appeler "la Toile".

ZIP
Fichiers obtenus après compression par le logiciel PKzip (extension .ZIP).

ANNEXE II : **Noms de domaine et glossaire des Acronymes et institutions représentatives de l'Internet**

Extension classiques (TLDs Top-Level Domains).

.com	Usage à caractère économique et commercial. [Entreprises et sociétés]
.org	Usage « réservé » aux organisations à caractère non commercial et aux associations à but non lucratif.
.net	Usage « réservé » aux organisations offrant des services Internet ou de téléphonie à une très grande échelle.
.edu	Usage à caractère éducatif. [Écoles privées et publiques, lycées et universités]
.int	Réservé à un usage international.
.mil	Usage réservé aux forces militaires US
.gov	Usage réservé aux institutions gouvernementales US
.arpa	Propriété de l'ARPA (Advanced Research Project Agency).

Nouvelles Extension (TLDs Top-Level Domains).

.biz	[JV Team/Neustar.com, *USA*]
.info	[Afilias/Skadden Arps, *USA*]
.name	[The Global Name Registry Ltd, *UK*]
.eu	[Europe]
.aero	[SITA, *Genève*]
.coop	[National Coop. Business Association, *USA*]
.museum	[Swedish Museum & Getty Museum, *USA*]
.pro	[RegistryPro, Ltd/Hayes & Curran, *Ireland*]

Glossaire des Acronymes et institutions représentatives de l'Internet

Acronym	Explanation	More Information
ACP	Administrative Challenge Panels	http://www.gtld-mou.org/docs/tracps.htm http://www.gtld-mou.org/docs/dispute.html
ADR	Alternative Dispute Resolution	http://www.gtld-mou.org/docs/dispute.html
AFA	Association des Fournisseurs d'Acces à Internet (French Access Providers Association)	http://www.afa-france.com
AFNIC (NIC-France)	Association Française pour le Nommage Internet en Coopération	http://www.nic.fr
AfriNIC	African NIC	http://www.afrinic.org
AFTLD	African ccTLDs	http://www.wwtld.org/aftld.txt
AIRA	American Internet Registrants Association	http://www.aira.org
APRAM	Association des Practiciens en Droits des Marques et des Modèles	-
APTLD	Council of the Asia Pacific country code Top Level Domains	http://www.aptld.org
APNG	Asia Pacific Networking Group (APNG)	http://www.apng.org

APNIC	Asia-Pacific Network Information Center	http://www.apnic.net
ARIN	American Registry for Internet Numbers.	http://www.arin.net
ARPA	Advanced Research Projects Agency (See also DARPA)	http://www.darpa.mil
BIND	Berkeley Internet Name Domain	http://www.isc.org/bind.html
CABASE	Camara Argentina de Internet, el Comercio Electrónico, los Contenidos y Servicios On Line	http://www.cabase.org.ar
CAIP	Canadian Association of Internet Providers	http://www.caip.ca
CDT	Center for Democracy and Technology	http://www.cdt.org
CENTR	Council of European National Top level domain Registries	http://www.centr.org
CIGREF	Club Informatique des Grandes Entreprises Françaises	http://www.cigref.fr
CIX	Commercial Internet Exchange	http://www.cix.org
CNRI	Corporation for National Research Initiatives	http://www.cnri.reston.va.us
CORE	Council of Registrars	http://www.corenic.org
CORE-MoU	Council of Registrars Memorandum of Understanding	http://www.gtld-mou.org/docs/core-mou.htm
DARPA	Defense Advanced Research Projects Agency	http://www.darpa.mil
DNS	Domain Name System	-
DOC	US Department of Commerce	http://www.doc.gov
EC	European Commission/European Community	http://www.eu.int
ECTA	European Communities Trade Mark Association	http://www.ecta.org
eCOMLAC	Federacion Latinoamericana y del Caribe para Internet y el Comercio Electronico	http://www.ecom-lac.org
ENRED	Foro Latinoamericano de Redes	http://www.reuna.cl/vi-foro/
ETNO	European Public Telecommunications Network Operators Association	http://www.etno.be
ETSI	European Telecommunications Standards Institute	http://www.etsi.fr http://www.etsi.org
EuroInternet	European Internet Business Association	http://www.eurointernet.org
EuroISPA	European Internet Services Provider Associations	http://www.euroispa.org
FCC	US Federal Communications Commission	http://www.fcc.gov
gTLD	generic Top Level Domain (not associated with country code)	

gTLD-MoU	Generic Top Level Domain Memorandum of Understanding	http://www.gtld-mou.org
IAB	Internet Architecture Board	http://www.iab.org/iab
IAHC	International Ad Hoc Committee	http://www.iahc.org
IANA	Internet Assigned Numbers Authority	http://www.iana.org
ICANN	Internet Corporation for Assigned Names and Numbers	http://www.icann.org
ICC	International Chamber of Commerce	http://www.iccwbo.org
IETF	Internet Engineering Task Force	http://www.ietf.org
IESG	Internet Engineering Steering Group	http://www.ietf.org/iesg.html
INTA	International Trademark Association	http://www.inta.org
IOPS.ORG	Group of commercial Internet Service Providers	http://www.iops.org
iPOC	gTLD-MoU Interim Policy Oversight Committee	http://www.gtld-mou.org
ISA	Interactive Services Association	http://www.isa.net
ISO	International Organization for Standardization	http://www.iso.ch
ISOC	Internet Society	http://www.isoc.org
ISP	Internet Service Provider	-
ISPA-uk	Internet Services Providers Association of the UK	http://www.ispa.org.uk/
ITAA	Information Technology Association of America	http://www.itaa.org
ITU	International Telecommunication Union	http://www.itu.int
LACTLD	Latin American and Caribbean ccTLDs	http://www.lactld.org
MARQUES	Association of European Brand Owners	http://www.martex.co.uk/marques/
MoU	Memorandum of Understanding	-
MPAA	Motion Picture Association of America	http://www.mpaa.org
NANC	North American Numbering Council	-
NANP	North American Numbering Plan	-
NIC	Network Information Center	-
NSF	US National Science Foundation	http://www.nsf.gov
NSI	Network Solutions, Inc.	http://www.netsol.com
NSI Registrar	The initial Registry-Registrars functions of the NSI has been splitted into two names: NSI Registry and NSI Registrar	http://www.netsol.com

VeriSign Global Registry Services	Sep 2000: the name "NSI Registry" has been changed to "VeriSign Global Registry Services"	http://www.netsol.com
NTIA	US National Telecommunications and Information Agency	http://www.ntia.doc.gov
OECD	Organization for Economic Co-operation and Development	http://www.oecd.org
PAB	gTLD-MoU Policy Advisory Body	http://www.gtld-mou.org
POC	gTLD-MoU Policy Oversight Committee	http://www.gtld-mou.org
RFC	Request for Comments	-
RIPE	Réseaux IP Européens	http://www.ripe.net
SRS	Shared Registratry System	http://www.gtld-mou.org/press/core-1.htm
TCP/IP	Transmission Control Protocol/Internet Protocol	-
TLD	Top Level Domain	-
URL	Uniform Resource Locator	-
USPTO	United States Patent and Trademark Office	http://www.uspto.gov
WIPO	World Intellectual Property Organization	http://www.wipo.int
WTO	World Trade Organization	http://www.wto.org
WWTLD	World-wide Alliance of Top Level Domains	http://www.wwtld.org

Pour en savoir plus, voici une liste de liens qui peuvent être utiles :

ICANN et la gouvernance de l'Internet

- **ICANN** - "Internet Corporation for Assigned Names and Numbers" http://www.icann.org/
- **IANA** - "Internet Assigned Numbers Authority" http://www.iana.org/
- **DNSO** - "Domain Name Supporting Organization" http://www.dnso.org/
- **ASO** - "Address Supporting Organization" http://www.aso.icann.org/
- **PSO** - "Protocol Supporting Organization" http://www.pso.icann.org/
- **GAC** - "Governmental Advisory Committee" http://www.noie.gov.au/projects/international/DNS/gac/index.htm

Organismes coordonnant l'adressage et le routage

- **RIPE** - "Reseaux IP Européens" http://www.ripe.net/

- **APNIC** - "Asie Pacific NIC"
  http://www.apnic.net/
- **ARIN** - "American Registry for Internet Numbers"
  http://www.arin.net/

Organismes de standardisation

- **IETF** - "Internet Engineering Task Force"
  http://www.ietf.org/
- **W3C** - "World Wide Web"
  http://www.w3c.org/
- **ITU** - "International Telecom Union"
  http://www.itu.org/
- **ETSI** - "European Telecommunications Standards Institute"
  http://www.etsi.org/
- **IAB** - "Internet Architecture Board"
  http://www.iab.org/
- **ISOC** - "Internet Society"
  http://www.isoc.org/

Organisme international

- **WIPO** - "Word Intellectual Property Organization""
  http://www.wipo.org/

Organismes européens

- **ISPO** - "Information Society Promotion Office - European Internet Forum"
  http://www.ispo.cec.br/eif/
- **EUROPA** - "European Union"
  http://www.europa.eu.int/

Organismes régionaux

- **CENTR** - "Council of European National Top level domain Registries"
  http://www.centr.org/
- **AFTLD** - "African Top Level Domains"
  http://www.aftld.org/
- **APTLD** - "Asia-Pacific Top Level Domain forum"
  http://www.aptld.org/
- **LACTLD** - "Latin American & Caribbean Country Code Top Level Organization"
  http://www.lactld.org/
- **NATLD** - "North American Top Level Domain Organization"
  http://www.natld.org/

Collèges du DNSO

- **collège "ccTLD Registries" du DNSO**
  http://www.wwtld.org/
- **collège "business" du DNSO**
  http://www.bcdnso.org/
- **collège "gTLD Registries" du DNSO**
  http://www.gtldregistries.org/
- **collège "ISP and connectivity Providers" du DNSO**
  http://www.dnso.org/constituency/ispcp/ispcp.html/
- **collège "Non commercial domain name holders" du DNSO**

http://www.ncdnhc.org/

- **collège "Registrars" du DNSO**
  http://www.dnso.org/constituency/registrars/registrars.html/
- **collège "Intellectual property" du DNSO**
  http://ipc.songbird.com/

Registre du ".ch"

- **SWITCH**
  http://www.nic.ch/

Registre de ".com" ".org" ".net"

- **InterNic**
  http://www.internic.net/

Liste officielle des administrateurs des 240 ccTLD (codes ISO ".de" ".fr" ...)

- **IANA** - Root zone
  http://www.iana.org/cctld/cctld-whois.htm

Autres Analyses

- http://www.icannwatch.org
- http://www.civilsocietyinternetforum.org
- http://www.cpsr.org/dns/index.html
- http://www.cdt.org/dns/icann/elections

Source : http://www.gouvernance-internet.com.fr/liens.html

ANNEXE III

TEXT MINING ET INTELLIGENCE ECONOMIQUE : AUJOURD'HUI ET DEMAIN

Xavier Polanco

Unité de Recherche et Innovation
Institut de l'Information Scientifique et Technique
Centre National de la Recherche Scientifique
polanco@inist.fr

La Fouille de Données Textuelles (FDT), c'est-à-dire le Text-Mining (TM), est ici présentée par rapport à l'Intelligence Economique (IE). L'*intelligence cycle* (Pinkerton, 1994) implique la conversion de l'information primaire (en anglais *raw information*) en information utile à l'entreprise. Dans la mesure où cette information primaire se trouve sous la forme de documents, de données textuelles, et qu'il s'agit de la transformer en connaissance, la FDT apparaît pour la *competitive intelligence* ou intelligence économique comme le moyen adéquat pour accomplir cette tâche essentielle. Ce texte présente d'abord ce que la FDT représente aujourd'hui, pour ensuite conclure avec quelques remarques sur ses perspectives.

INTRODUCTION

L'argumentation industrielle ou commerciale en faveur du *text mining* dans le contexte de l'intelligence économique est souvent basée sur le fait qu'un pourcentage assez important de l'information à traiter par les entreprises, en vue de la prise de décisions stratégiques, est de nature textuelle.

Le World Wide Web est encore une autre raison en faveur de l'intérêt pour le *text mining*. En effet, avec le Web les données non structurées (telles que le texte) sont devenues le type prédominant de données en ligne. Dans ce cadre, l'information utile ne se trouve pas être explicite comme dans une base des données de type relationnel, mais implicite au sens où elle est « enfouie » dans les textes, d'où la métaphore de la « fouille » (ou en anglais *mining*) : le système doit extraire l'information qui a été encodée dans le texte par son auteur.

Souvent la veille technologique et l'intelligence économique sont présentées comme des activités connexes ou bien similaires sinon synonymes. Ce fait permet d'étendre l'apport de la fouille de données textuelles au domaine de la veille technologique et scientifique, dans la mesure où l'information scientifique et technique est de nature textuelle tels que les articles scientifiques, la documentation technique et les brevets.

PRESENTATION

Cette présentation s'adresse aux praticiens de la veille et de l'intelligence économique. Son objectif est de montrer ce que la fouille des données textuelles représente. Elle peut également être étendue aux praticiens des études quantitatives de la science et de la technologie, notamment à ceux pour qui la science est analysée au travers des publications et la technologie au travers des brevets. Les publications scientifiques et les brevets sont des données textuelles dont s'occupe justement le *text mining*.

Plan :

1. Data Mining et Text Mining
2. Architecture et système
3. Techniques et méthodes
4. Traitement linguistique
5. Structure de classification
6. Extraction de règles d'association

Chaque item de ce plan de présentation sera développé avec le souci de fournir une information synthétique sans chercher à développer une véritable argumentation technique. L'ambition est de fournir l'information nécessaire pour se faire une vision de la fouille de données textuelles.

1 - Data Mining et Text Mining

Commençons par la distinction entre *data mining* et *text mining* c'est-à-dire entre fouille des données et fouille de données textuelles.

Le but de la fouille de données a été définie comme "the non trivial extraction of implicit, previously unknown, and potentially useful information from given data" (Frawley et al, 1991, p. 1-27, cité in Feldman, 1998, p. 65). Ou encore : "The non trivial process of identifying valid, novel, potentially useful, and ultimately understable patterns in data" (Fayyad et al., 1999).

Historiquement, le *data mining* est à la base du *text mining* au sens où celui-ci est l'extension du même but et du même processus vers des données textuelles. La distinction est donc fondée à son origine principalement sur la nature des données auxquelles s'adressent l'une et l'autre, d'une part des données numériques et factuelles, et d'autre part des données textuelles. Un autre élément de distinction est l'état de structuration des données. En général le *data mining* travaille sur des données structurées et stockées dans des bases de données

relationnelles. En revanche, le *text mining* travaille sur des données textuelles non structurées (Feldman et al., 1998a et 1998b; Landau et al., 1998).

Le *text mining* se distingue du *data mining* également par les moyens techniques spécifiques qu'il le faut employer pour traiter les données textuelles et non structurées.

Une définition générale du *text mining* est la suivante : l'extraction d'information à partir des formes ou patrons non manifestes (au sens de *hidden patterns*) dans des grands corpus de textes. Autrement dit, l'objectif est le traitement de grandes quantités d'information qui sont disponibles sous une forme textuelle et non structurée. (Feldman et al., 1998a ; Landau et al., 1998).

L'intelligence économique est sensée assurer aux acteurs économiques une information exploitable et utile, dans la mesure où cette information est textuelle (notes, lettres, rapports techniques, articles scientifiques, brevets, etc.), l'intérêt que le *text mining* peut représenter pour l'intelligence économique vis-à-vis du simple *data mining* est alors évident.

2 – Architecture et système

Considérons maintenant les principaux outils composants d'un système de fouille de textes en général. Ici on se limite à l'esquisse d'une architecture générale et abstraite. L'important est de savoir que chacun de ces outils est indispensable pour mener à bien une opération de fouille de données textuelles.

1. Un outil d'accès et collecte des données
2. Un outil d'ingénierie du document
3. Un outil d'ingénierie du langage écrit
4. Un outil de fouille (ou *mining tool*)
5. Un outil de visualisation

La théorie, les méthodes et les techniques appliquées à l'occasion de la conception et du développement de chacun de ces cinq outils, leur donnant ainsi une réalité technologique particulière, ce sont des éléments à tenir en compte. On voit donc que les systèmes de fouille de données textuelles sont des systèmes complexes réunissant des compétences diverses.

L'outil d'accès et de collecte des données textuelles doit être capable d'opérer aussi bien à partir du Web sur de documents HTML, que sur des bases de données soit bibliographiques soit textuelles au sens du texte plein (ou *full-text*). L'outil d'ingénierie du document sert à la gestion et le traitement des documents qui sont sous la forme de données hétérogènes et sans structure fixe, dites données semi-structurées (DSS), afin de leur appliquer un formalisme du type SGML ou XML et réaliser ainsi l'étiquetage de leurs attributs (par exemple, la date, le titre, les auteurs, la source, le corps du texte, et l'ensemble de termes caractérisant le document). L'outil d'ingénierie linguistique est destiné au traitement du langage écrit pour l'extraction de termes et l'indexation automatique des documents, mais aussi pour la gestion

de ressources terminologiques telles que thesaurus, vocabulaires d'indexation, bases de termes, etc. L'outil de fouille de textes réalise la fonction générale de l'acquisition de connaissances à partir des données textuelles qui ont été collectées, formatées et indexées préalablement. Et enfin, l'outil de visualisation doit fournir à l'utilisateur les moyens hypertexte et graphiques pour explorer et analyser les résultats.

Les propriétés interactives et itératives de cette architecture s'avèrent être d'une extrême importance dans la fouille de textes comme dans tout autre système destiné à l'analyse de l'information, où l'intervention des experts du domaine est nécessaire, et les retours en arrière dans le processus sont des décisions fréquentes. D'où la nécessité d'une architecture informatique modulaire et flexible du système.

Cette présentation va par la suite se concentrer (ci-dessous sections 4, 5 et 6) sur la nature des outils [3] et surtout [4], parce qu'ils représentent les éléments les plus spécifiques d'un système de fouille des données textuelles.

Quant aux produits commerciaux de fouille de données ou *data mining*, une remarque générale est qu'ils ne sont au fond que des outils statistiques (Ultsch, 1999) : "The terms Data Mining and Knowledge Discovery are often used in those systems in an inflationary way for statistical tools enhanced with a fancy visualization interface. The difference between exploratory statistical analysis and Data Mining lies in the aim which is sought. Data Mining aims at Knowledge Discovery". Un système de *data mining* ne se réduit pas à un simple outil d'analyse statistique des données. L'intention finale de la fouille des données est donc l'extraction de connaissance (en anglais *knowledge discovery*).

3 – Techniques et méthodes très diverses

Selon l'appel au *Text Mining Workshop de l'International Joint Conference on Artificial Intelligence (IJCAI 99)*, http://ijcai.org/, les techniques utilisées par le *text mining* sont celles de la fouille des données (ou *data mining*), de l'apprentissage automatique, de la recherche d'information, de la compréhension du langage naturel, du raisonnement à partir de cas, des statistiques, et enfin de la gestion de connaissances. Le but étant d'aider les personnes à obtenir de la connaissance à partir de grandes quantités de textes semi-structurés.

Comme le montre *Text Mining Workshop IJCAI 99*, la fouille de données textuelles est un terme recouvrant des activités très diverses. Selon Toussaint, Simon et Cherfi (2000), une première différence entre les méthodes vient des données qui sont fournies à l'algorithme de fouille et de la qualité de ces données selon la capacité des algorithmes à prendre en compte des données de qualité "inférieure". Le second aspect de différenciation porte sur l'algorithme de fouille, sur le type de données qui sont fournies à l'utilisateur final, qui, dans tous les cas, doit être un expert.

Nous avons vu ci-dessus, dans la section 1, que le *texte mining* se distingue du *data mining* par les moyens techniques qu'il le faut employer pour traiter les données textuelles. Ces données sont des textes et aussi des données non structurées ou semi-structurées. De là donc deux tâches : traiter automatiquement le langage naturel dans sa forme écrite, manipuler des

données non structurées ou semi-structurées. Lesquelles demandent des outils spécialement adaptés.

Au sujet du problème de la manipulation de données semi-structurée (DSS), Al Hulou, Napoli et Nauer (2000) analysent comment le langage de description de documents XML, avec les outils qui lui sont associés et l'essor qu'il connaît, peut servir comme un formalisme de représentation intermédiaire entre DSS et représentation de connaissances par objet (RCO). Comme il a été dit plus haut, l'intention finale de la fouille de données textuelles est l'extraction de connaissances, d'où le besoin également d'un système de représentation de connaissances et de raisonnement (être capable de faire des inférences).

4 - Traitement linguistique

La capacité à traiter automatiquement le langage écrit apparaît comme une étape importante de la fouille de données textuelles. La plupart des systèmes ont relayé au second plan les données issues de l'indexation manuelle et exploitent les résultats d'une indexation automatique.

L'approche d'ingénierie linguistique est la suivante. En entrée des données textuelles que l'on doit soumettre à un traitement permettant l'extraction automatique d'éléments linguistique plus complexes que des simples mots. L'étiquetage des textes (ou *tagging*), l'assignation automatique de catégories morpho-syntaxiques telles que le nom, le verbe, l'adjectif, etc., aux mots du document, et la lemmatisation, sont les étapes de ce traitement. Ensuite vient la phase de l'extraction de termes à partir des textes étiquetés, laquelle est suivie d'une phase de filtrage. Ce filtrage est généralement statistique et il consiste en calculer un score aux termes. Les termes sont sélectionnés en fonction de leur score. Seulement les termes ayant un score supérieur à un seuil déterminé sont sélectionnés comme candidats pour l'indexation de documents.

L'indexation des documents peut se faire avec les termes que l'on obtient soit par une extraction fondée sur de patrons syntaxiques, soit à partir d'un référentiel terminologique, tel qu'un thesaurus, et de méta-règles de variation. Toussaint, Simon et Cherfi (2000) utilisent cette dernière méthode. Feldman et ses collègues (1998b) utilisent la première approche.

Les expériences prouvent que l'approche linguistique assure une meilleure performance des algorithmes de fouille. Dans l'article "Text Mining at the Term Level", Feldman et ses collègues (1998b) montrent l'intérêt de travailler au niveau du terme et non du mot. Ainsi ils désignent leur système comme un "term-based text mining system".

La capacité à manipuler de données semi-structurées, l'exploitation d'une indexation automatique fondée sur une analyse morphologique et syntaxique des textes sont des conditions préalables et nécessaires mais pas suffisantes. Pour que la fouille à proprement parler se réalise, il faut encore l'application d'algorithmes capables de construire une structure classificatoire (taxonomie) et d'effectuer l'extraction de règles d'association

Passons donc à ce que l'on peut considérer comme le cœur du processus de la fouille de données textuelles.

5 – Structure de classification

La nécessité d'une taxonomie est une question cruciale pour la fouille de textes. La taxonomie est construite dans le but de structurer l'ensemble de termes hiérarchiquement. Une telle structure classificatoire est importante pour la plupart d'algorithmes de fouille de textes. Le système doit donc disposer d'un moyen de construction de la taxonomie en question.

Ainsi par exemple Simon (2000) montre que la théorie des treillis de Galois permet de produire à la fois un outil de classification hiérarchique et un outil de construction de règles d'association. Toussaint, Simon et Cherfi (2000) proposent une méthode de fouille de données fondée sur les treilles de Galois et sur l'extraction de règles d'association en vue d'aider des experts dans leur tâche de veille scientifique. Rappelons au passage que les treilles de Galois sont connus aussi sous l'appellation de *conceptual clustering*. Les treillis de Galois opèrent avec les notions d'intension et d'extension et la relation de subsomption. Un treillis de Galois permet la construction des deux types de structures propres à la fouille de données textuelles : [1] une structure de classification qui regroupe les documents en fonction des termes qui leurs sont associés et réciproquement ; [2] l'extraction de règles d'association entre les termes associés aux documents.

Quelle qu'elle soit la méthode de construction de cette taxonomie, il est important de noter que chaque nœud représente un concept. Dans le cas d'une taxonomie fondée sur le treillis de Galois : chaque élément du treillis est considéré comme un concept formel et le graphe (diagramme de Hasse) comme une relation de généralisation/spécialisation entre les concepts. Le treillis est donc perçu comme une hiérarchie de concepts. Chaque concept est une paire composée d'une extension représentant un sous-ensemble des instances de l'application et d'une intention représentant les propriétés communes aux instances (Godin et al. 1995).

L'aspect pragmatique de la taxonomie. Elle permet à l'utilisateur de définir les tâches de fouille d'une manière concise. Ceci suppose une interface de visualisation graphique et de navigation dans la structure classificatoire (taxonomie) et les règles d'association obtenues et d'observer le type de relation existant entre les termes participant à une règle.

Un exemple (Feldman et al., 1998) : "the user can specify interest only in the relationships of *companies* in the context of *business alliances*. In order to do so, we need two nodes in the term taxonomy marked *business alliances* and *companies*. The first node contains all terms related to alliance such as *joint venture, strategic alliance, combined initiative* etc., while the second node is the parent of all *company* names".

La construction de cette structure classificatoire permet de mettre en évidence les concepts potentiellement intéressants pour l'analyste. De plus, elle permet l'extraction de règles d'association.

6 - Extraction de règles d'association

Les règles d'association ont été présentées en 1993 par R. Agrawal, T. Imielinski et A. Swani dans leur article "Mining Association Rules between Sets of Items in Large Databases". La signification intuitive d'une règle d'association X = > Y, où X et Y sont des ensembles d'items, est qu'une transaction contenant X est susceptible de contenir également Y (Agrawal et al. 1996). L'application type est l'analyse des données du panier de supermarché, où des règles, comme celle-ci, par exemple, "34% de tous les clients qui achètent de poissons également achètent du vin blanc", peuvent être trouvées. Les règles d'association s'avèrent par ailleurs être tout à fait utiles dans des applications économiques.

Les règles d'association peuvent être calculées soit par l'algorithme d'Agrawal, comme c'est le cas dans Feldman (1998b) ; soit à partir des treillis de Galois comme le propose Simon (2000) et le font Toussaint, Simon, Cherfi (2000). Ce second approche est tout récent et il est encore au niveau de la recherche (au sein de l'équipe Orpailleur du LORIA à Nancy) - http://www.loria.fr

Les règles d'association extraient des patrons à partir des données du type [jus de raisin => chromatographie] : celle-ci montre que dans le corpus analysé, les documents s'intéressant au jus de raisin le font systématiquement en rapport avec la chromatographie ; [histamine => amine biogène] : l'histamine est une amine biogène qui est tout particulièrement étudiée dans le corpus par sa toxicité dans les aliments.

Lors de la phase d'interprétation, il est indispensable de disposer d'un outil de visualisation et navigation.

PERSPECTIVES

La nouvelle économie et avec elle la gestion croissante de connaissances dans la vie des organisations sont des facteurs définissant un nouvel horizon pour la veille et l'intelligence économique, mais aussi pour la bibliométrie qui les est associée. Dans ce nouveau contexte, la demande de fouille de données textuelles de la part de la veille et de l'intelligence économique ne peut que s'accroître. Si cette demande se développe, elle devra en exercer un effet d'orientation sur la recherche dans le domaine de la fouille de données textuelles et sur la mise au point de systèmes viables. Ceci pose le problème de savoir quel est l'état de l'offre du côté de la fouille de données textuelles.

Notre présentation a voulu montrer succinctement ce que représente un système de fouille de données textuelles (section 2), la diversité de disciplines et méthodes que la fouille de données textuelles mobilise (section 3), et puis l'état actuel de la fouille de données textuelles sur le plan de l'ingénierie linguistique (section 4) et de l'ingénierie de la connaissance (section 5 et 6). Pour l'analyse des perspectives de la fouille de données textuelles, il est nécessaire de tenir en compte les sections 2 et 3, autrement dit le fait de son appartenance à un réseau

multidisciplinaire et dans lequel elle devra évoluer suivant une fertilisation croisée. Tenir compte également de ce que nous avons évoqué dans les sections 4, 5 et 6 reconnaissant les deux dernières comme le noyau propre ou strictement spécifique de la fouille de données textuelles.

L'évolution de la fouille de données textuelles est ainsi liée à un ensemble de disciplines informatiques dont le souci principal est de savoir comment traiter à l'aide des ordinateurs les contenus de l'information et leur conversion en connaissances. Admettons de les grouper sous le label commun de *technologies de l'intelligence*. Ces domaines de recherche seraient les suivants :

- Extraction d'information (Cowie et Lehnert, 1996).
- Traitement automatique du langage naturel.
- Visualisation de l'information (Card et al., 1999).
- Recherche d'information mais dans sa nouvelle de version de *modern information retrieval* (Baeza-Yates et Ribeiro-Neto, 1999).
- Gestion de connaissances (O'Leary, 1998).

Ce cluster de recherches constitue le voisinage de la fouille des données textuelles, laquelle appartient en propre au cluster noyau formé par :

- Fouille de données (*Data Mining*)
- Fouille de données de la Toile (*Web Mining* ou *Internet Data Mining*)
- Fouille de données textuelles (*Text Mining*)
- Extraction de connaissances (*Knowledge Discovery in Databases*)

Dans ce réseau scientifique et technologique, l'avenir de la relation entre la fouille des textes et la veille et l'intelligence économique est en train de se construire.

AVERTISSEMENT

Ce document est exclusivement destiné aux participants du colloque *Veille technologique, Intelligence économique et Bibliométrie*. Colloque organisé par la section Bibliothéconomie du DEC Sciences du Livre de l'Université Catholique de Louvain-la-Neuve, les 23-24 janvier 2001. Seulement après l'intervention, la discussion et les remarques dans le cadre de ce colloque, l'auteur envisage de le transformer dans un article pour être publié

BIBLIOGRAPHIE

R . Agrawal, H. Mannila, R. Srikant, H. Toiven, A. Ikeri Verkamo (1996) Fast Discovery of Association Rules, in Fayyad et al. (1996) p. p.307-328.

R. Al Hulou, A. Napoli, E. Nauer (2000) XML : un formalisme de représentation intermédiaire entre donnée semi-structurées et représentations par objets, in C. Dony, H. A. Sahraoui (eds) *Langages et Modèles à Objets*. Paris, HERMES, p. 75-90.
R. Baeza-Yates, B. Ribeiro-Neto (1999) *Modern Information Retrieval*. ACM Press / Addison-Wesley Longman.
S. K. Card, J. D. MacKinlay, B. Schneiderman (eds) (1999) *Readings in Information Visualization. Using Vision to Think.* San Francisco, Cal., Morgan Kaufman Publishers, Inc.
J. Cowie, W. Lehnert (1996) Information Extraction, *Communications of the ACM*, vol. 30 (1), p. 80-91.
U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds) (1996) *Advances in Knowledge Discovery and Data Mining.* Menlo Park, Cal. AAAI Press / The MIT Press.
U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth (1999) Data Mining and Knowledge Discovery in Databases: Introduction to the Special Issue. *Communications of the ACM*, vol. 39 (1).
R. Feldman, Y Aumann, A. Zilberstein, Y. Ben-Yuda (1998a) Trend Graphs: Visualizing the Evolution of Concept Relationships in Large Document Collections, in Zytkow et Quafafou (1998) p. 38-46.
R. Feldman, M. Fresko, Y. K Kinar, Y Lindell, O. Liphstar, M. Rajman, Y. Scheler, O. Zamir (1998b) Text Mining at the Term Level, in Zytkow et Quafafou (1998) p. 65-73.
R. Godin, G. Mineau, R. Missaoui (1995) Méthodes de classification conceptuelle basées sur les treillis de Galois. *Revue d'intelligence artificielle*, vol. 9 (2), p. 105-137.
D. Landau, R. Feldman, Y. Aumann, M. Fresko, Y. Lindell, O. Lipshtat, O. Zamir (1998) TextViz: An Integrated Visual Environment for Text Mining, in Zytkow et Quafafou (1998) p. 56-64.
D. E. O'Leary (1998) Knowledge Management Systems: Converting and Connection. *IEEE Intelligent Systems*, vol. 1 (3), p. 30-33.
R. L. Pinkerton (1994) Competitive Intelligence Revisited: A History and Assessment of Its Use in Marketing. *Competitive Intelligence Review*, vol. 5 (4), p. 23-31.
A. Simon (2000) *Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données.* Thèse de doctorat de l'Université Henri Poincaré – Nancy 1.
Y. Toussaint, A. Simon, H. Cherfi (2000) Apport de la fouille de données textuelles pour l'analyse de l'information. *Actes des Journée Francophones d'Ingénierie des Connaissances* (IC'2000), Toulouse, p. 335-344.
A. Ultsch (1999) Data Mining and Knowledge Discovery with Emergent Self-organizing Feature Map for Multivariate Time Series, in E. Oja, S. Kaski (eds) *Kohonen Maps*. Amsterdam, ELSEVIER, p. 33-45.
J. M. Zytkow, et M. Quafafou (eds) (1998) *Principles of Data Mining and Knowledge Discovery*. Proceedings of the Second European Symposium, PKDD'98, Nantes. Berlin, Springer, (Lecture Notes in Artificial Intelligence 1510).

ANNEXE 1V : **La répartition géo-stratégique de l'Internet**

Map of the Root Servers

Carte de la répartition des serveurs racines à travers le monde. 80 à 90% du traffique IP transite par ces serveur dont 10 sur 13 sontt situés au Etats-Unis

name	org	city	type
a	NSI	Herndon, VA, US	com
b	USC-ISI	Marina del Rey, CA, US	edu
c	PSInet	Herndon, VA, US	com
d	U of Maryland	College Park, MD, US	edu
e	NASA	Mt View, CA, US	usg
f	Internet Software C.	Palo Alto, CA, US	com
g	DISA	Vienna, VA, US	usg
h	ARL	Aberdeen, MD, US	usg
i	NORDUnet	Stockholm, SE	int
j	NSI (TBD)	Herndon, VA, US	(com)
k	RIPE	London, UK	int
l	ICANN	Marina del Rey, CA, US	org
m	WIDE	Tokyo, JP	edu

Source : http://www.icann.org/correspondence/roberts-testimony-14feb01.htm